UNIT 1

FUNDAMENTALS

Look for the Logo on the Title Cover Before You Buy
SIA GROUP
SIA GROUP

Syllabus

FUNDAMENTALS

Need for Low Power Circuit Design, Sources of Power Dissipation – Switching Power Dissipation, Short Circuit Power Dissipation, Leakage Power Dissipation, Glitching Power Dissipation, Short Channel Effects –Drain Induced Barrier Lowering and Punch Through, Surface Scattering, Velocity Saturation, Impact Ionization, Hot Electron Effect.

LEARNING OBJECTIVES

- Need of Low power circuit design
- Switching power dissipation and Short circuit power dissipation in CMOS circuits
- Expression for short circuit power dissipation in CMOS logic gates
- Leakage power dissipation, Reverse diode leakage current and Sub threshold leakage current
- Glitch power dissipation, and concept of short channel effects
- Drain-induced barrier lowering, punch through, velocity saturation, surface scattering phenomena in short-channel effect
- Impact ionization and hot electron effect

INTRODUCTION

Power consumption has been a prominent challenge in VLSI design due to decreasing feature size and increasing chip density and operation frequency. The IC's high power consumption hinders their usage in portable systems. It also causes overheating, lowering performance and chip life. The need for portable communication and computer devices has raised the need to optimize chip power consumption. Therefore, low-power design is a vital technology in today's semiconductor industry.

PART-A SHORT QUESTIONS WITH SOLUTIONS

Q1. What is the need of Low Power VLSI?

Ans: Model Paper-1, Q1(a)

Power consumption has been a prominent challenge in VLSI design due to decreasing feature size and increasing chip density and operation frequency. The IC's high power consumption hinders their usage in portable systems. It also causes overheating, lowering performance and chip life. The need for portable communication and computer devices has raised the need to optimize chip power consumption. Therefore, low-power design is a vital technology in today's semiconductor industry.

Q2. What are the sources of power dissipation in conventional CMOS digital circuits?

Ans:

In standard CMOS digital circuits, the average power dissipation can be divided into three types, which are as follows:

1. Dynamic (switching) power dissipation,
2. Short-circuit power dissipation, and
3. Leakage power dissipation.
4. Glitching power dissipation. This is a special case of dynamic power dissipation.

Q3. What is Glitch and Glitch power dissipation?

Ans: Model Paper-1, Q1(f)

Glitch

A glitch is any unwanted pulse at the output of a digital circuit. In other words, a glitch is a small spike that happens at the output of a digital circuit before the output reaches its steady state.

Glitch Power Dissipation

Glitching power dissipation is the power consumed by the glitches. In other words, glitch power dissipation is the power dissipated in the intermediate transitions during the evaluation of the logic function of the circuit.

Q4. What are two most significant leakage current components observed in a MOSFET?

Ans: Model Paper-2, Q1(f)

The two most significant leakage current components observed in a MOSFET transistor are,

1. **Reverse Diode leakage Current**

Reverse diode leakage current is one of the most signifioant leakage current components that contribute significantly to leakage power dissipation. Whenever the pn-junction between the drain and the bulk of the transistor is reversely biased, reverse diode leakage occurs. In other words, reverse saturation current is extracted from the power supply by the reverse-biased drain junction.

2. **Sub Threshold Leakage Current**

Sub-threshold current is a type of leakage current that arises in CMOS circuits. This current is produced by the diffusion of carriers between the source and drain regions of the transistor when it is in a weak inversion.

Q5. What are short channel effects? List five different short channel effects.

Ans: Model Paper-3, Q1(e)

Short-channel effects are a set of events that occur when the MOSFET's channel length gets closer to the width of the space charge regions of the source and drain junctions.

List of five different short channel effects are

1. Drain-induced barrier lowering and punch through
2. Surface scattering
3. Velocity saturation
4. Impact ionization
5. Hot electron effect.

Q6. What are adverse effects caused by hot caused by hot electrons effect.

Ans: Model Paper-2, Q1(a)

A few of the adverse affects associated with hot carrier effects on MOSFET behavior encompass: A shift in threshold voltage, reduction in transconductance, additional leakage currents, instabilities and excess noise.

Q7. What are the methods to reduce punch through effect?

Ans: Model Paper-3, Q1(a)

The punch through effect can be reduced by increasing the degree of doping in the transistor body (bulk). The drain and source depletion regions would be minimized in this case, and no parasitic current path will form. Other methods to reduce the effect of punch through are, use of spatially constrained dopant implantations, such as

(a) Halo or pocket implantations and

(b) Delta doping.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

PART-B ESSAY QUESTIONS WITH SOLUTIONS

1.1 Need for Low Power Circuit Design

Q8. What is the need of Low power circuit design? Explain.

Ans:

There has been a lot of progress in low-power design over the last few years to keep power consumption (and heat dissipation) to a minimum in very-high density VLSI chips. The main driving force behind the development of high-density VLSI chips is portable devices such as smart phones, tablets, flash memory devices, handheld mobile telephones, personal digital assistants (PDAs), and laptop computers. Because of the ever-decreasing feature size and increase in chip density, power consumption has become a big issue in high-density VLSI design. The use of integrated circuits in portable systems is constrained due to their excessive power dissipation. Portable devices need very low power consumption (and dissipation) along with high chip density and high throughput. Thus, low-power digital integrated circuit design has evolved as a highly active and quickly emerging field of CMOS design.

In high-performance digital systems, such as microprocessors, digital signal processors (DSPs), and other applications, the chips have high clock frequencies. If the clock frequency of the chip increases, then the power dissipation of the chip, and thus the temperature, will increase linearly. To control the temperature levels, the chips need specialized and costly packaging and cooling arrangements, which would result in a further escalation of the system cost. An alternative solution to this issue is the use of a low-power VLSI design.

Another problem that indicates the need for low-power design is the reliability of VLSI chips. There is indeed a significant connection between the peak power dissipation of digital circuits and the reliability of VLSI chips, such as electron migration and hot-carrier induced device degradation. Further, the thermal stress generated by heat dissipation on the chip is the main reliability problem. Therefore, to enhance the reliability of VLSI chips, their power consumption is reduced using low-power design techniques.

1.2 Sources of Power Dissipation-Switching Power Dissipation, Short Circuit Power Dissipation, Leakage Power Dissipation, Glitching Power Dissipation

Q9. Explain switching power dissipation in CMOS circuits.

Ans: Model Paper-3, Q3(b)

Switching power dissipation is also known as dynamic power dissipation. The switching power dissipation occurs whenever there is a switching in the output of the CMOS logic gate. In other words, power dissipation occurs when the output node of a CMOS logic gate makes a transition (HIGH-LOW or LOW-HIGH). CMOS circuits generate dynamic power dissipation by the charging and discharging of capacitances.

Consider the circuit illustrated in Figure (1) that demonstrates how dynamic power dissipation occurs during switching. In this circuit, a two-input NOR gate is used to operate two NAND gates, which are connected together by interconnection lines. The total capacitive load at the output of the NOR gate is composed of three components:

1. The output capacitance of the NOR gate itself,
2. The total interconnect capacitance, and
3. The input capacitances of the gates that are being driven.

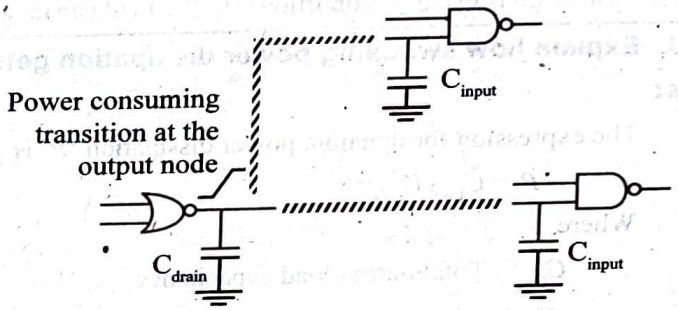


Figure (1): NOR Gate Driving NAND Gates

The output capacitance of the NOR gate denoted as C_{drain} is primarily made up of junction parasitic capacitances, which are formed by the drain diffusion areas of the MOS transistors in the circuit and contribute to the overall capacitance of the NOR gate. The interconnect lines that connect the gates together make contributions to the total interconnect capacitance and is denoted as $C_{interconnect}$. The input capacitances denoted as C_{input} are primarily caused by the gate oxide capacitances of the transistors that are linked to the input terminal.

A CMOS Logic Gate in Generic Form

As illustrated in figure (2), a CMOS logic gate that undergoes an output transition can be described by the nMOS network, pMOS network, and the total load capacitance linked to its output node. Figure (2) illustrates the generic (basic) form of a CMOS logic gate.

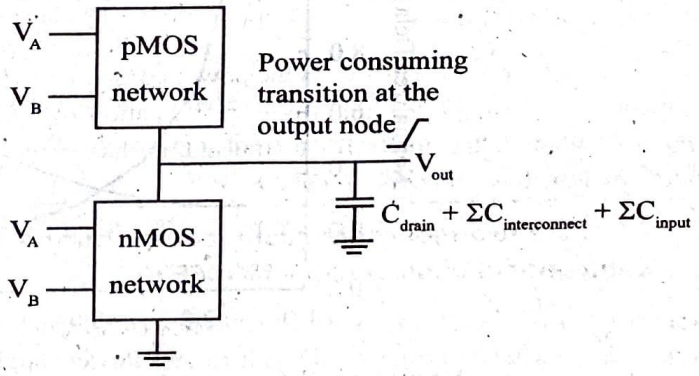


Figure (2): Generic form of a CMOS Gate

The average dynamic power dissipation of the CMOS logic gate, driven by a periodic waveform with preferably zero rise- and fall-times, can be determined by calculating the energy needed to charge up the output node (V_{out}) to V_{DD} and charge down the total output load capacitance to ground. Mathematically it is expressed as,

$$P_{avg} = \frac{1}{T}\left[\int_0^{T/2} V_{out}\left(-C_{load}\frac{dV_{out}}{dt}\right)dt + \int_{T/2}^{T}(V_{DD}-V_{out})\left(C_{load}\frac{dV_{out}}{dt}\right)dt\right] \quad \ldots (1)$$

The equation for the average dynamic power consumption in CMOS logic circuits is obtained by simplifying the integral in equation (1).

$$P_{avg} = \frac{1}{T}C_{load}V_{DD}^2 \quad \ldots (2)$$

$$P_{avg} = C_{load}\cdot V_{DD}^2 \cdot f \quad \ldots (3)$$

From equation (3), it is clear that, the average dynamic power dissipation is related to the square of the power supply voltage V_{DD}. So, lowering the voltage of the power source will reduce power consumption significantly. The average dynamic power dissipation is exactly proportional to the load capacitance C_{load} hence lowering C_{load} will also decrease power dissipation.

Q10. Explain how switching power dissipation gets affected by reducing the power supply voltage V_{DD}.

Ans:

The expression for dynamic power dissipation 'P' is given as,

$$P = C_{load}\cdot V_{DD}^2 \cdot f \quad \ldots (1)$$

Where,

C_{load} – Total output load capacitance

V_{DD} – Power supply voltage

f – Frequency of transitions at the output

From equation (1), it is clear that, the average dynamic power dissipation is related to the square of the power supply voltage V_{DD}. So, lowering the voltage of the power source will reduce power consumption significantly. However, the inevitable design trade-off is an increase in the propagation delay. This is demonstrated for the CMOS inverter circuit by inspecting the propagation delay formula shown below.

$$\tau_{PHL} = \frac{C_{load}}{k_n(V_{DD}-V_{T,n})}\left[\frac{2V_{T,n}}{V_{DD}-V_{T,n}} + \ln\left(\frac{4(V_{DD}-V_{T,n})}{V_{DD}} - 1\right)\right] \quad \ldots (2)$$

If the power supply voltage is reduced while keeping the other variables constant, then from equation (2) it is clear that the propagation delay will increase. The graph in figure (1) illustrates the normalized variation of the propagation delay as a function of power supply voltage V_{DD}. In this example, the threshold voltages of the nMOS and pMOS transistors are considered to be constant with $V_{T,n}$ and $V_{T,p}$ being equal to 0.08V and – 0.8V, respectively. Similarly, the normalized variation in dynamic power dissipation as a function of the supply voltage V_{DD} is shown in the same graph.

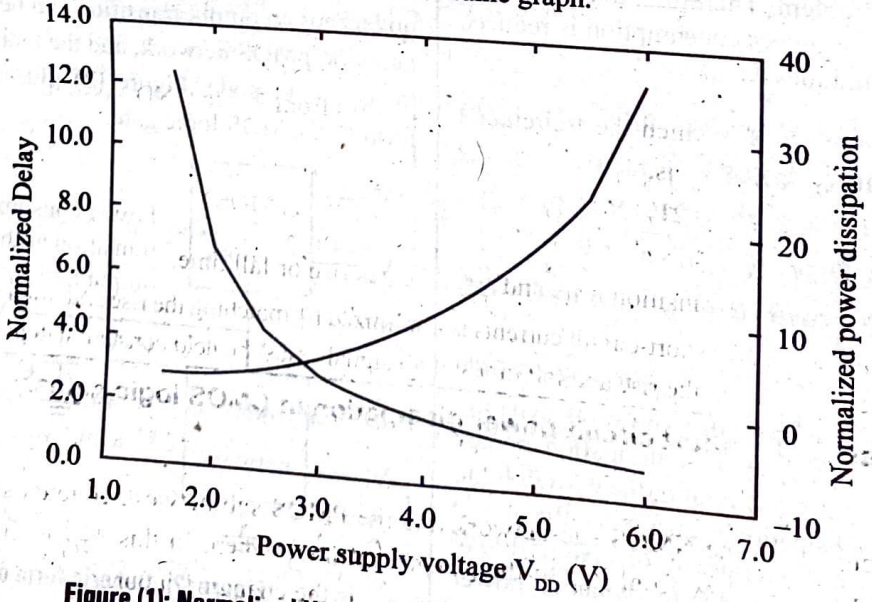


Figure (1): Normalized Variation of the Delay as a Function of V_{DD}

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

The study of dynamic power dissipation described above is based on the premise that the output node of a CMOS gate undergoes only one power-consuming transition (0-to-V_{DD}) in every single clock cycle. But, this assumption may not always be right. Depending on the circuit design, logic style, and input signal characteristics, the node transition rate might be lower than the clock rate.

In order to properly describe this behavior, we shall introduce the concept of α_T (node transition factor), which represents the effective number of power-consuming transitions occurring in every single clock cycle. Then, the expression for average dynamic power dissipation is given as

$$P_{avg} = \alpha_T C_{load} V_{DD}^2 f_{CLK} \quad \text{... (3)}$$

Q11. Explain short circuit power dissipation in CMOS logic gates.

Model Paper-1, Q2(a)

Ans:

Usually, for a CMOS logic gate (with input waveforms having zero rise and fall times), when the pull-up network (PMOS transistors) is turned OFF, the pull-down network (NMOS transistors) is turned ON and vice versa. Both the PUN and PDN are not conducted simultaneously. But, in practice, the input waveforms driving the CMOS logic gate are associated with finite rise and fall times. As a consequence, both the pull-up and pull-down networks in the CMOS gate conduct simultaneously for a short period of time during the finite slope period of inputs. This creates a direct current path between the power supply and ground, as illustrated in figure (1). The direct current path is also known as short circuit current path and the power dissipation caused by this is known as short circuit power dissipation.

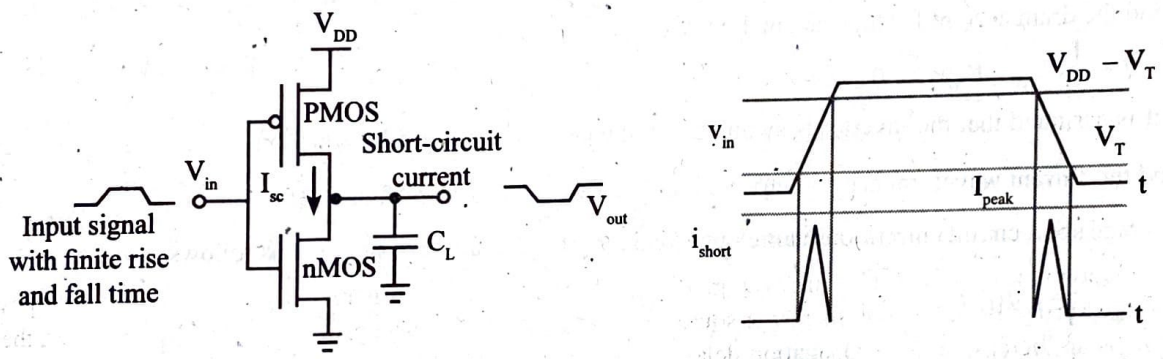


Figure (1): Short-circuit Current Flowing Through Both NMOS and PMOS Transistors Simultaneously

The short circuit current is highly dependent on the output load capacitance as well as the rise and fall durations of the input signals. When the output load capacitance is smaller and the input signals rise and fall durations are longer, the short circuit current will be quite high. With a larger output load capacitance and reduced rise and fall duration of input signal, the short circuit current will be extremely low.

When a short circuit current flows through a CMOS logic gate, the energy consumed (E) per switching period and the average power consumption (P) are given by the following equations

$$E = V_{DD}\frac{I_{peak}t_{sc}}{2} + V_{DD}\frac{I_{peak}t_{sc}}{2} = t_{sc}V_{DD}I_{peak} \quad \text{... (1)}$$

$$P = t_{sc}V_{DD}I_{peak}f = C_{sc}V_{DD}^2 f \quad \text{... (2)}$$

Here, 't_{sc}' represents the time for which short-circuit current flows from V_{DD} to GND. Assuming the rise and fall durations are linear, a close estimate of this time 't_{sc}' is,

$$t_{sc} = \frac{V_{DD} - 2V_T}{V_{DD}} t_s \approx \frac{V_{DD} - 2V_T}{V_{DD}} \times \frac{t_{r(f)}}{0.8} \quad \text{... (3)}$$

Where, 't_s' denotes the 0-100% transition time and $t_{r(f)}$ indicates rise or fall time.

The power consumption caused by short-circuit currents is minimized by matching the rise and fall times of the input and output signals. In other words, at the circuit level, the rise and fall times of all signals must be held constant within a fixed range at all times.

Q12. Derive the expression for short circuit power dissipation in CMOS logic gates.

Ans:

Model Paper-2, Q2

In static CMOS circuits both the NMOS sub-network and the PMOS sub-network conduct simultaneously and a short circuit path arises for direct current flow from V_{DD} to the ground terminal, when. In this V_{TH_n} is threshold voltage of NMOS transistors, V_{TH_p} is the threshold voltage of PMOS transistors and V_{IN} is the changing input.

The direct current flow (from V_{DD} to the ground terminal) produces short circuit power dissipation and its average val is given as $I_{avg}V_{DD}$. To obtain the expression of average short circuit power dissipation we consider a symmetrical inverter. F a symmetrical inverter illustrated in figure (1), we have $\beta = \beta_n = \beta_p$ and $V_T = V_{T,n} = |V_{T,p}|$ this makes the analysis simple. Furth the input signal applied to the inverter is also symmetrical that is the signals rise time = fall time as illustrated in figure (2) a is periodic with period T.

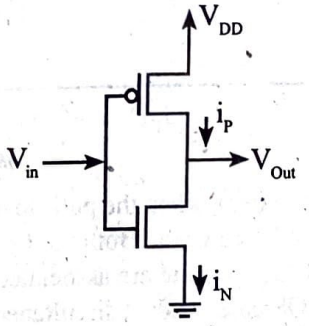


Figure (1): Symmetrical Inverter

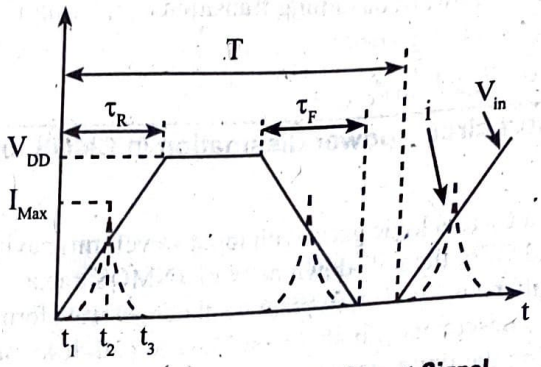


Figure (2): Symmetrical Input Signal

Short-circuit Power Dissipation

From figure (2), it is evident that the short circuit current increases from 0 to I_{max} during the internal t_1 to t_2. The NMOS is in saturation and the drain current for long channel transistor is given as,

$$I = \frac{\beta}{2}(V_{In} - V_{TH})^2 \text{ for } 0 \le T \le I_{max} \quad \ldots (1)$$

Since it is assumed that the inverter is symmetrical, the current expressed by equation (1) reaches its peak value when $V_{IN} = \frac{V_{DD}}{2}$ and the current waveform is symmetrical about the vertical axis $t = t_2$ (figure (2)).

The average short circuit current is obtained by integrating the instantaneous current as follows,

$$I_{avg} = \frac{1}{T}\int_0^T I(t)\,dt \quad \ldots (2)$$

Substituting equation (1) in equation (2), we get,

$$I_{avg} = 2\frac{2}{T}\int_{t_1}^{t_2} \frac{\beta}{2}(V_{In}(t) - V_{TH})^2 dt \quad \ldots (3)$$

Since the raising and falling portions of the input waveform are linear ramps, we have,

$$V_{in}(t) = \frac{V_{DD}}{\tau}t \quad \ldots (4)$$

At time t_1, the transistor starts conducting. Hence, $V_{In}(t_1) = V_{TH}$ and $V_{In}(t_1) = \frac{V_{DD}}{\tau}t_1$

$$\Rightarrow \quad t_1 = \frac{V_{in}(t_1)}{V_{DD}}\tau = \frac{V_{TH}}{V_{DD}}\tau \quad \ldots (5)$$

At time t_2, the current is maximum (peak) and $V_{in}(t) = \frac{V_{DD}}{2}$.

$$V_{In}(t_2) = \frac{V_{DD}}{\tau}t_2$$

$$\Rightarrow \quad t_2 = \frac{V_{In}(t_2)}{V_{DD}}\tau = \frac{V_{DD}}{2V_{DD}}\tau = \frac{\tau}{2} \quad \ldots (6)$$

Substituting equations (4), (5) and (6) into equation (3), we get,

$$I_{avg} = 2\frac{2}{T}\int_{\left(\frac{V_{TH}}{V_{DD}}\right)\tau}^{\tau/2} \frac{\beta}{2}\left(\frac{V_{DD}t}{\tau} - V_{TH}\right)^2 dt \quad \ldots (7)$$

Let $\frac{V_{DD}}{\tau}t - V_{TH} = \theta$, then $dt = \frac{d\theta\tau}{V_{DD}}$ ... (8)

and $\theta_1 = 0$ and $\theta_2 = \left(\frac{V_{DD}}{2} - V_{TH}\right)$ ... (9)

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Substituting equation (8) and equation (9) into equation (7), we get,

... (10)

$$I_{avg} = \frac{2\beta}{T}\int_0^{\left(\frac{V_{DD}}{2}-V_{TH}\right)} \theta^2 d\theta \frac{\tau}{V_{DD}}$$

$$\Rightarrow \quad I_{avg} = \frac{2\beta\tau}{TV_{DD}}\int_0^{\left(\frac{V_{DD}}{2}-V_{TH}\right)} \theta^2 d\theta$$

$$\Rightarrow \quad I_{avg} = \frac{2\beta\tau}{TV_{DD}}\left[\frac{\theta^3}{3}\right]_0^{\frac{V_{DD}}{2}-V_{TH}}$$

$$\Rightarrow \quad I_{avg} = \frac{2\beta\tau}{TV_{DD}}\left[\frac{\left(\frac{V_{DD}}{2}-V_{TH}\right)^3}{3} - 0\right]$$

$$\Rightarrow \quad I_{avg} = \frac{2\beta\tau}{TV_{DD}}\frac{(V_{DD}-2V_{TH})^3}{24}$$

Therefore, the expression for average short circuit current is,

... (11)

$$I_{avg} = \frac{\beta}{12T}(V_{DD}-2V_{TH})^3 \frac{\tau}{V_{DD}}$$

Therefore, the expression for average short circuit power dissipation P_{sc}

... (12)

$$P_{sc} = I_{avg} V_{DD}$$

$$\therefore P_{sc} = \frac{\beta}{12}(V_{DD}-2V_{TH})^3 \frac{\tau}{T}$$

Q13. Write short notes on leakage power dissipation. Model Paper-1, Q3(b)

Ans:

The nMOS and pMOS transistors included in every CMOS logic gate often exhibit non-zero reverse leakage and sub-threshold currents. In a CMOS VLSI chip (that encompasses an enormous number of transistors), these currents contribute significantly to the total power dissipation even though the transistors are not undergoing any switching activity. Leakage power dissipation is the term used to describe the dissipation that occurs as a result of leakage currents. The leakage current magnitude is primarily affected by processing parameters.

The two most significant leakage current components observed in a MOSFET transistor are

1. Reverse diode leakage current
2. Sub threshold leakage current.

Q14. Explain Reverse diode leakage current. Model Paper-3, Q2(a)

Ans:

Reverse Diode Leakage Current

Reverse diode leakage current is one of the most significant leakage current components that contribute significantly to leakage power dissipation. Whenever the pn-junction between the drain and the bulk of the transistor is reversely biased, reverse diode leakage occurs. In other words, reverse saturation current is extracted from the power supply by the reverse-biased drain junction. For instance, take a CMOS inverter (figure 1) with a high input voltage, by which it turns OFF the PMOS transistor and turns ON the nMOS transistor and discharges the output node voltage to zero (making the output LOW). Even if the pMOS transistor is turned off, the reverse voltage difference of VDD between its drain and the n-well will cause diode leakage through its drain junction. In addition, the n-well section of the pMOS transistor is reverse-biased relative to the p-type substrate. This will generate another considerable leakage current component.

When the input to the CMOS inverter is low (equal to zero), it turns OFF the nMOS transistor, turns ON the pMOS transistor, and charges the output node voltage to VDD (making the output HIGH). Under these circumstances, a condition similar to the one described above may be seen. Even though the nMOS transistor is turned off, the reverse voltage difference between its drain and the p-type substrate will generate a reverse leakage current, which is also pulled from the power supply (through the pMOS transistor).

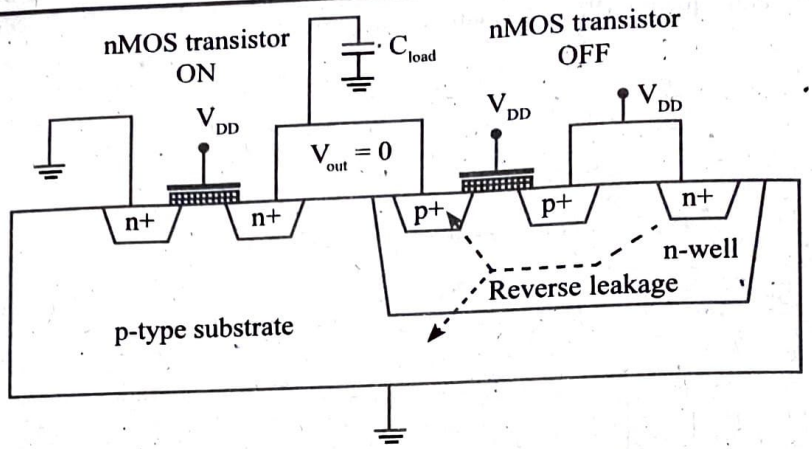


Figure (1): Reverse Leakage Current in CMOS Circuit

The magnitude of the reverse leakage current of a pn-junction is evaluated using the following equation,

$$I_{reverse} = AJ_S\left(e^{\frac{qV_{bias}}{kT}} - 1\right) \quad \ldots(1)$$

In this equation, V_{bias} represents the magnitude of the reverse bias voltage across the junction, J_S represents the reverse saturation current density and A is the area of the junction. Usually, the magnitude of the reverse saturation current density increases remarkably with temperature. In essence, the pn-junction reverse bias leakage current is a function of junction area and doping concentration and strongly associates with temperature.

Q15. Explain Sub threshold leakage current.

Ans: Model Paper-1, Q2(a)

Sub-threshold current is a type of leakage current that arises in CMOS circuits. This current is produced by the diffusion of carriers between the source and drain regions of the transistor when it is in a weak inversion. When the MOS transistor is operating in the sub-threshold region, its sub-threshold current exhibits an exponential relationship with the gate voltage. The magnitude of the sub-threshold current becomes considerable (large) when the gate-to-source voltage is lesser than and extremely close to the threshold voltage of the device. Especially in this particular situation, the magnitude of power dissipation caused by the sub-threshold leakage current can be equivalent to the switching (dynamic) power dissipation of the circuit.

Figure (1) illustrates the sub-threshold leakage current in the CMOS circuit.

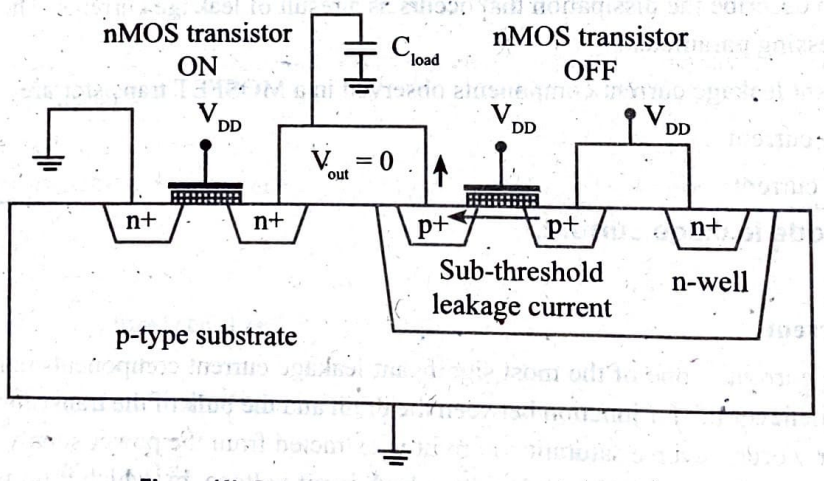


Figure (1): Sub-threshold Current in CMOS Circuit

The sub-threshold current component is very crucial in the evaluation of the total power dissipation when the circuit is in standby operating mode. This is because sub-threshold leakage current occurs even though there is no switching activity in the circuit. The sub-threshold current of a MOS transistor is exponentially related to its terminal voltages and is given as

$$I_{D(subthreshold)} \cong \frac{q_{D_n} W x_c n_0}{L_B} \cdot e^{\frac{q\phi_r}{kT}} \cdot e^{\frac{q}{kT}(A \cdot V_{GS} + BV_{DS})} \quad \ldots(1)$$

To keep the sub-threshold current component under control, one relatively simple measure is to avoid using very low threshold voltages. This way, the V_{GS} of the nMOS transistor can be kept safely below $V_{T,n}$ when the input is logic zero and the $|V_{GS}|$ of the pMOS transistor can be kept safely below $|V_{T,p}|$ when the input is logic one.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Q16. Explain Glitch power dissipation.

Ans:

As per definition, a glitch is any unwanted pulse at the output of a digital circuit. In other words, a glitch is a small spike that happens at the output of a digital circuit before the output reaches its steady state. Glitches are produced in a circuit due to the difference in signal arrival times at the inputs of a gate. For instance, consider the digital circuit illustrated in figure (1). Each gate in this circuit has a unit delay. From figure (1), it is clear that there is a difference in the arrival times of signals at the inputs of the AND gate. Because of this, there is a glitch of 1 unit width in the output of the AND gate. This means before the output of the AND gate reaches a steady state, its input has multiple transitions. As a result, there is a glitch in its output. Glitches occur mainly due to a difference or inequality in the path lengths in the logic network. When all of a gate's input signals arrive at the same time, there is no glitching.

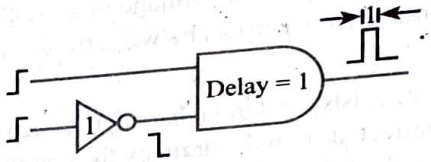


Figure (1)

Glitching power dissipation is the power consumed by the glitches. In other words, glitch power dissipation is the power dissipated in the intermediate transitions during the evaluation of the logic function of the circuit. The power consumed is directly proportional to the number of transitions at the input of a gate; these unnecessary transitions increase power consumption. Glitch power usually amounts to about 20% of the overall power consumption of the chip, and even 70% in some typical cases, such as the combinational adder. For instance, consider the digital circuit illustrated in figure (2). The OR gate produces a delay of 1 unit. The transient response of the output contains three edges: two rising and one falling. This is a glitch with a combined width of 2 units. The number of edges in the transient response at the output of a gate is equal to the number of arriving signals at the gate's input (number of transitions). The maximum difference in arrival time of signals at the inputs of a gate is called "differential path delay" and is also the maximum width of the possible glitch at the circuit output.

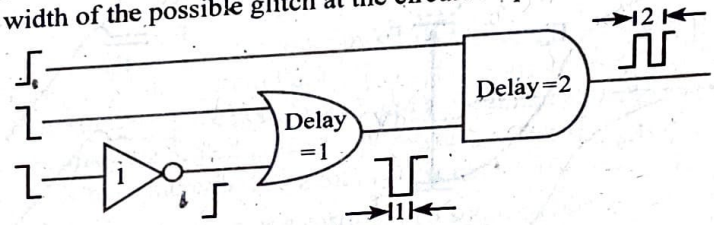


Figure (2)

One of the earliest methods to reduce glitch power was path balancing. Due to differences in the arrival times of signals at gate inputs, glitches are produced. The concept behind this approach is to avoid glitch generation by balancing the delays of the paths so that at any specified gate in the digital circuit, the signals arrive at its input at the same time. This is described with the help of the digital circuits illustrated in figures (3) and (4). In figure (3), the signals arrive at the gates at different times, causing glitches. Restructuring the circuit, as seen in Figure (4), prevents the occurrence of glitches.

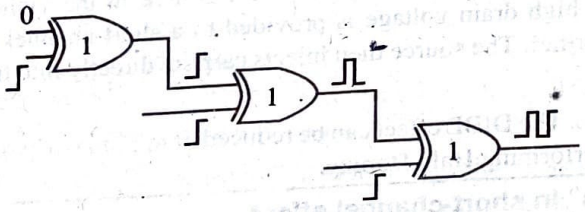


Figure (3)

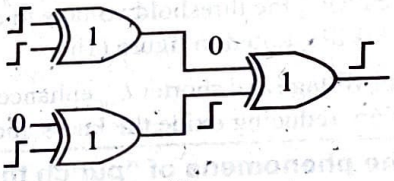


Figure (4)

1.3 Short Channel Effects – Drain Induced Barrier Lowering and Punch Through, Surface Scattering, Velocity Saturation, Impact Ionization, Hot Electron Effect

Q17. What are short channel effects? Explain in brief about the problems caused by short channel effects.

Model Paper-2, Q3(b)

Ans:

Short-channel effects are a set of events that occur when the MOSFET's channel length gets closer to the width of the space charge regions of the source and drain junctions. Under this condition, a limitation is imposed on electron drift characteristics in the channel and threshold voltage is modified due to the shortening channel length. As a consequence, these produce multiple problems such as polysilicon gate depletion, threshold voltage roll-off, drain-induced barrier lowering (DIBL), velocity saturation, reverse leakage current increase, mobility decrease, hot carrier effects, and other disturbances.

The reduced threshold voltage makes it harder to totally switch off the transistor. Electrostatic interaction betwe source and drain causes the gate to become inefficient due to the DIBL effect. The current drive is reduced by velocity satura The power dissipation is increased by the leakage current. Increased surface scattering reduces charge carrier mobility, lowe output current. Aside from these considerations, impact ionization and hot carrier effects damage MOSFET performance, caus the transistor to behave differently from long-channel devices.

Q18. Explain the "drain-induced barrier lowering" short-channel effect.

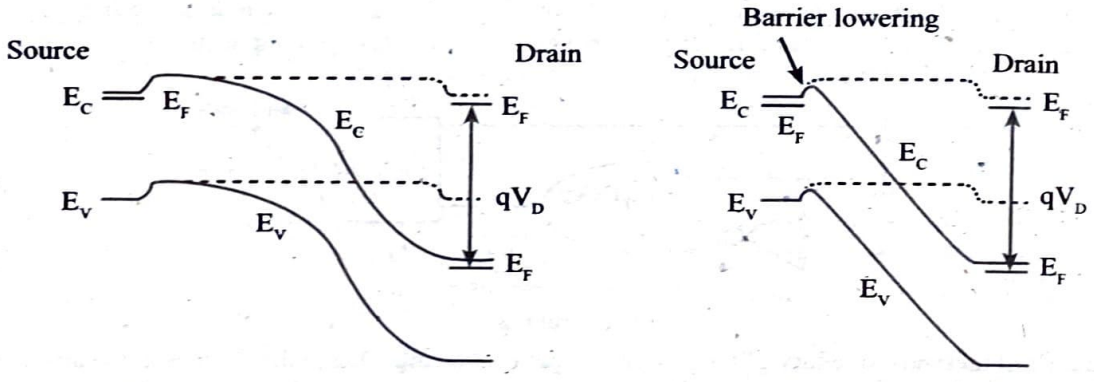
Ans:

Model Paper-I, Q

Drain-induced barrier lowering (DIBL) is a short-channel phenomenon occurring in MOSFETs. It reduces the transis threshold voltage with increasing drain voltages. The threshold voltage of a MOSFET is the least gate-to-source voltage nece to generate an inversion region for conduction of charge carriers between the drain and the source, when the drain and source properly biased.

In a MOSFET with a long channel, there exists a potential barrier between source and drain which is lowered by applyi gate voltage. The barrier is electrostatically protected from the drain by the conjunction of the substrate and gate. The source a drain are separated far enough in long-channel devices that their depletion areas have little influence on the potential barrier a hence the threshold voltage is almost independent of channel length and drain voltage.

This is no longer true in short-channel devices. In short channel devices in addition to the gate voltage, drain voltag also has a significant effect on reducing this barrier. In this drain is pretty close to control the channel formation, so a high dra voltage can open the barrier and switch on the transistor prematurely. The DIBL effect not only reduces threshold voltage it al increases sub-threshold current. The effect is illustrated with the help of energy band diagram in figure (1).



(a) Long Channel MOSFET **(a) Short Channel MOSFET**

Figure (1)

DIBL arises when the depletion layer of the drain comes into contact with the source in the vicinity of the channe surface to decrease the source potential barrier. When a high drain voltage is provided to a short-channel device, the barrie height is reduced, causing the threshold voltage to drop further. The source then injects carriers directly into the channel surface bypassing the gate as illustrated in figure (1b).

Higher drain voltage and shorter L_{eff} enhances DIBL. The DIBL effect can be reduced or minimized by increasing substrate doping concentration, reducing oxide thickness and by performing Halo Doping.

Q19. Explain the phenomena of "punch through" in short-channel effect.

Ans:

"Punch through" is a term used to describe a phenomenon occurring in MOSFET short-channel devices. When the depletion regions of the drain and source junctions meet and create a single depletion region in MOSFETs, punch through is said to occur. In other words, because the channel length is diminished while keeping the doping constant, the distance between both the drain and source depletion region boundaries is reduced. The voltage V_{DS} acts as a reverse bias for these depletion regions and an increase in V_{DS} causes the borders to be pushed farther away from the junction and closer to each other. Since the channel length is reduced, the reverse bias voltage V_{DS} provokes the depletion regions to merge, then it is said that a punch-through has occurred. Under this condition, the current flow in the channel is not regulated by the gate voltage, but instead it increases swiftly with an increase in the drain-source voltage V_{DS} (even if the gate voltage is zero, current continues to flow). Punch through current is mostly determined by the applied drain voltage and the depth of the source/drain junction.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

When the reverse bias voltage V_{DS} is increased above the value needed to trigger the punch-through, it decreases the potential barrier level for the majority of carriers in the source. An increase in the drain voltage above the value needed to trigger the punch-through decreases the potential barrier for the majority of the carriers in the source. As a result, many of these carriers are able to cross the energy barrier and penetrate the substrate. The drain accumulates a few of them. The overall result is that the sub-threshold current increases. Moreover, punch-through provokes the sub-threshold slope to diminish. The parameter widely used to characterize the punch-through effect is the punch-through voltage V_{pT}. This parameter predicts the value of V_{DS} at which punch-through takes place with gate voltage $V_g = 0$. It is estimated approximately as the value of the V_{ds}, for which the total width of drain and source depletion region is equivalent to the effective channel length as follows:

$$V_{PT} \propto N_B(L - W_j)^3 \quad \text{... (1)}$$

Where,

N_B – Doping concentration at the bulk

L – Channel length

W_j – Junction width

The punch through effect may be reduced by increasing the degree of doping in the transistor body (bulk). The drain and source depletion regions would be minimized in this case, and no parasitic current path will form. Other methods to reduce the effect of punch through are use of spatially constrained dopant implantations, such as

(a) Halo or pocket implantations and

(b) Delta doping.

Q20. Explain in brief about velocity saturation in short-channel effect.

Model Paper-1, Q3(a)

Ans:

Velocity Saturation

When the size of the MOS device is reduced for low power VLSI design, the electric fields within the device increase significantly. In an effort to keep the electric field within limits, the voltages used in short-channel MOSFETs are lowered from 5 V to about 1 V. However, the channel lengths have gone from 5 μm to 70 nm, and the gate oxide thickness has gone from 100 nm to 1.5 nm. This clearly shows that sizes are scaled down proportionally faster than electric field reduction. Obviously, both the horizontal and vertical electric fields have increased by a lot in short-channel MOSFETs.

At higher electric fields, the drift velocity of charge carriers in a semiconductor will never rise linearly with the electric field. In contrast, the velocity of charge carriers begins to level out and then reaches its saturation (maximum) value V_{sat}. At room temperature, the value of saturation velocity V_{sat} for carriers in silicon is roughly 1×10^7 cm/sec. A range of models have been used to calculate the drift velocity as a function of the electric field E. One of the most widely used is,

$$v = \frac{\mu_n E_y}{\left\{1 + \left(\frac{E_y}{E_c}\right)^n\right\}^{1/n}} \quad \text{... (1)}$$

Where,

v is drift velocity,

μ_n is the low-field electron mobility,

E_c ($= V_{sat}/\mu_n$) is the critical field determining velocity saturation, and

E_y is the electric field in the direction of motion of the electrons.

When applied to low fields, equation (1) accurately produces $v = \mu_n E_y$ and when used for high fields, it correctly produces $v = v_{sat}$. The value of n is usually between 1 and 2 and is chosen to closely reflect experimental results.

The value of V_{Dsat} is essentially zero for relatively small values of L. This is comprehensible because it implies that the channel's velocity is saturated even at low drain voltages. The drain current I_D is then always saturated, and is approximately equal to,

$$I_{Dsat} = WC_{ox}v_{sat}[V_{GS} - V_T] \quad \text{... (2)}$$

From equation (2) it is evident that the drain current is independent of channel length L. That is, regardless of how small L is made, the current does not grow beyond a level as velocity saturation limits it. This means that decreasing L will not increase performance (greater currents and speeds) in future technologies.

Q21. Explain surface scattering phenomena occurring in short-channel effect.

Ans:

Model Paper-2, Q2(a)

Surface scattering: The vertical component of the electric field E_x accelerates electrons toward the surface, causing surface scattering. Due to surface scattering, mobility is reduced. The average surface mobility is nearly half that of the bulk mobility.

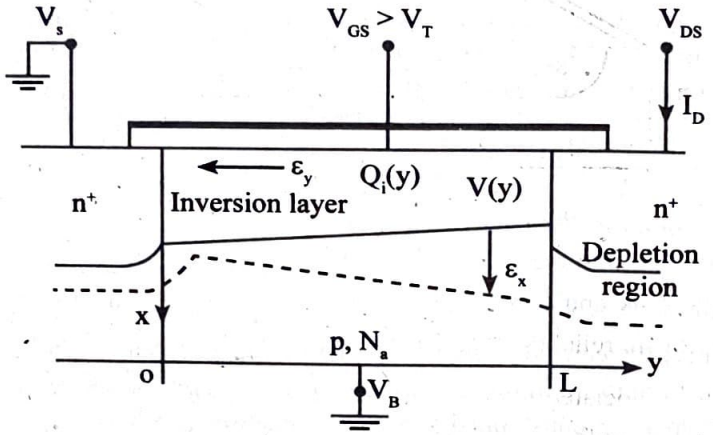


Figure (1)

The following is the physical explanation why transverse fields cause scattering and effect lateral current flow. In MOSFET, carriers travel via a thin inversion layer nearer the silicon surface (figure (1)), where they are subjected to numerous scattering events with the surface, especially when it is microscopically "rough". When the channel length becomes smaller because of the expansion of the depletion layer into the channel region, the longitudinal electric field component E_y increases, and the surface mobility becomes field-dependent. Because the carrier transport is confined within the narrow inversion layer, and the surface scattering (that is the collisions suffered by the electrons that are accelerated toward the interface by E_x) causes reduction of the mobility, the electrons move with great difficulty parallel to the interface, so that the average surface mobility, even for small values of E_y, is about half of the bulk mobility. Experiments have demonstrated that the mobility of electrons and holes in silicon can be expressed mathematically as,

$$\mu = \frac{\mu_0}{1 + (\theta E_{eff})^{\nu}} \quad \ldots (1)$$

Where,

E_{eff} is an effective transverse electric field,

μ_0 is the mobility without the transverse field effect, and

ν is a factor between 1 and 2.

The above equation has a physical origin, and can be derived using Mathiessen's rule. If $\mu 0$ represent mobility of carriers due to low transverse fields and μ_T represent mobility of extra scattering caused by the vertical field. Presuming that the latter scattering enhances with effective field, we have,

$$\frac{1}{\mu} = \frac{1}{\mu_0} + \frac{1}{\mu_r} = \frac{1}{\mu_0} + KE_{eff}^{\nu} \quad \ldots (2)$$

This gives us the μ_T equation we want.

Q22. Explain about impact ionization and hot electron effect.

Ans:

Model Paper-3, Q3(a)

The issue of hot electrons has continued to increase as technologies scale down because the dimensions of the devices are scaled down proportionally faster than voltage. This results in higher field strengths and thinner gate oxides in the devices.

The High fields cause electrons flowing through the channel to gain enough energy to become "hot". The hot carriers can cause impact ionization, that is, they strike the silicon atoms (in the drain) and ionize them. A consequence of this is the generation of extra holes dislodged from the drain. This causes an increase in drain current as well as a substrate current, that's made up of ionization-generated holes that flow towards the substrate (the transistor's most negative point). Some of the holes produced by impact ionization can also get heated up and be injected into insulator.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Another effect of hot carriers is that they may gain sufficient energy to overcome the potential barrier at the silicon-insulator interface, and get injected into the insulator. In other words, the hot electrons enter the gate-oxide layer and will produce gate current. A schematic of these effects is illustrated in figure (1).

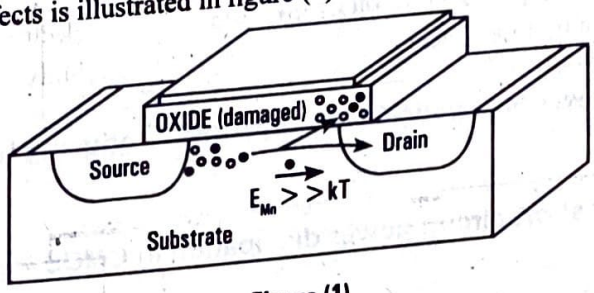


Figure (1)

The holes and electrons entering into the oxide layer cause many issues, including electron and hole trapping, interface state generation, and generation of bulk and "border" traps in the oxide layer. These phenomena are known as "hot carrier effects," and they constitute a significant reliability challenge for MOSFETs.

A few of the adverse affects associated with hot carrier effects on MOSFET behavior encompass: A shift in threshold voltage, reduction in transconductance, additional leakage currents, instabilities and excess noise.

IMPORTANT QUESTIONS

Q1. Explain switching power dissipation in CMOS circuits.

Ans: Refer Q9.

Important Ques

Q2. Explain how switching power dissipation gets affected by reducing the power supply voltage V_{DD}.

Ans: Refer Q10.

Important Ques

Q3. Derive the expression for short circuit power dissipation in CMOS logic gates.

Ans: Refer Q12.

Important Ques

Q4. Explain Reverse diode leakage current.

Ans: Refer Q14.

Important Ques

Q5. Explain Glitch power dissipation.

Ans: Refer Q16.

Important Ques

Q6. Explain the "drain-induced barrier lowering" short-channel effect.

Ans: Refer Q18.

Important Ques

Q7. Explain the phenomena of "punch through" in short-channel effect.

Ans: Refer Q19.

Important Ques

Q8. Explain in brief about velocity saturation in short-channel effect.

Ans: Refer Q20.

Important Ques

Q9. Explain surface scattering phenomena occurring in short-channel effect.

Ans: Refer Q21.

Important Ques

Q10. Explain about impact ionization and hot electron effect.

Ans: Refer Q22.

Important Ques

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

UNIT 2

Low Power Design Approaches and Switched Capacitance Minimization Approaches

Look for the Logo on the Title Cover Before You Buy
SIA GROUP
SIA GROUP

Syllabus

LOW POWER DESIGN APPROACHES

Low Power Design through Voltage Scaling – VTCMOS circuits, MTCMOS circuits, Architectural Level Approach – Pipelining and Parallel Processing Approaches.

SWITCHED CAPACITANCE MINIMIZATION APPROACHES

System Level Measures, Circuit Level Measures and Mask level Measures.

Learning Objectives

- Low power design through voltage scaling
- Operation of VTCMOS and MTCMOS circuits
- Concepts of pipe lining and parallel processing approaches
- Switched capacitances minimization at system level, circuit level and mask level

Introduction

The requirement of low-power design must be met along with equally demanding goals of chip density and high throughput. It is a major issue in high performance digital systems. The methods which are used to achieve low power consumption in this unit at circuit or transistor level. We discuss system level considerations such as pipe lining and hard ware replication (parallel processing).

PART-A SHORT QUESTIONS WITH SOLUTIONS

Q1. List out the observations on switching power reduction.

Ans:

The switching power consumption is expressed as,

$$P_{avg} = \left(\sum_{i=1}^{No.of\ nodes} \alpha_{Ti} . C_i . V_i \right) . V_{DD} . f_{clk}$$

On observing the equation, the average power consumption can be reduced by reducing,

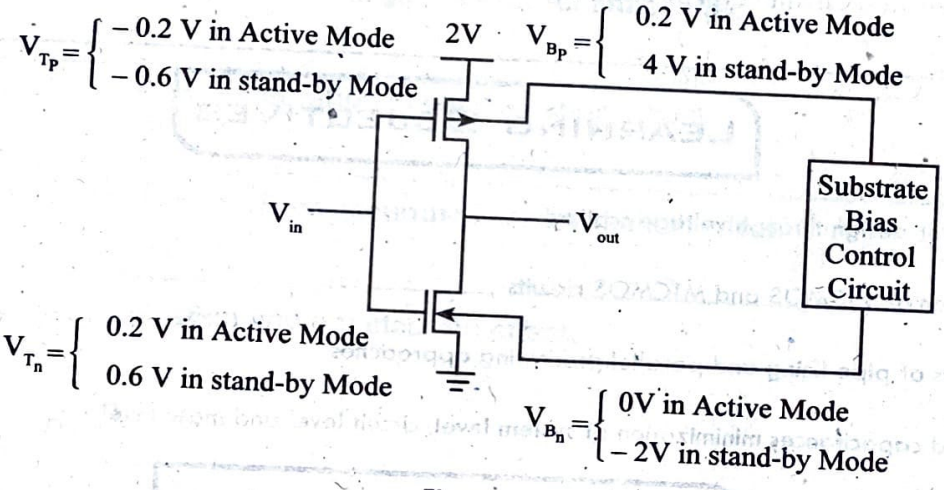
1. Power supply voltage, V_{DD}
2. Voltage switching in internal and output nodes
3. Node transition factor
4. Load capacitance.

Q2. Draw the circuit diagram of VTCMOS inverter.

Ans:

Model Paper-1, Q1(g)

The circuit diagram of a variable threshold CMOS inverter is as shown in figure,


V_{T_P} = { − 0.2 V in Active Mode; − 0.6 V in stand-by Mode }
2V
V_{B_P} = { 0.2 V in Active Mode; 4 V in stand-by Mode }
V_{in}
V_{out}
Substrate Bias Control Circuit
V_{T_n} = { 0.2 V in Active Mode; 0.6 V in stand-by Mode }
V_{B_n} = { 0V in Active Mode; − 2V in stand-by Mode }

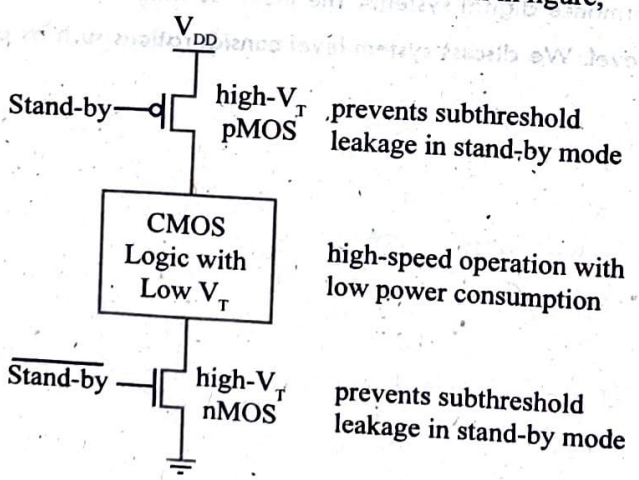
Figure

Q3. Draw the general structure of MTCMOS.

Ans:

Model Paper-2, Q1(g)

The circuit diagram of multiple-threshold CMOS inverter is as shown in figure,



Figure

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Q4. Give the delay expressions of a CMOS inverter?

Ans:

The delay expressions of a CMOS invereter are,

$$\tau_{PHL} = \frac{c_{load}}{k_n(V_{DD} - V_{T,n})}\left[\frac{2V_{T,n}}{V_{DD} - V_{T,n}} + \ln\left(\frac{4(V_{DD} - V_{T,n})}{V_{DD}} - 1\right)\right] \quad \ldots (1)$$

$$\tau_{PLH} = \frac{C_{load}}{k_n(V_{DD} - V_{T,n})}\left[\frac{2V_{T,p}}{V_{DD} - |V_{T,p}|} + \ln\left(\frac{4(V_{DD} - |V_{T,p}|)}{V_{DD}} - 1\right)\right] \quad \ldots (2)$$

Q5. When MTCMOS technique in used?

Model Paper-1, Q1(b)

Ans:

Multiple Threshold CMOS (MTCMOS) is the technique used to decrease the subthreshold current leakages and power dissipation in the circuit operating in stand-by mode.

Q6. What is pipeling approach?

Model Paper-2, Q1(b)

Ans:

The pipelining approach is a method of applying a data stream in a regular manner to a processing circuit in order to produce results in a sequential order. The supply voltage can be lowered with the pipelining approach, which saves power.

Q7. What is parallel processing approach?

Model Paper-3, Q1(b)

Ans:

- Parallel processing approach, like pipelining, can be used to reduce power dissipation by lowering the supply voltage. It trades-off area for lower power dissipation. It is also known as hardware replication.
- Parallel processing is the use of multiple copies of the same circuit to perform the same computation on multiple data sets simultaneously.

Q8. What is meant by switching activity?

Ans:

The switching activity refers to switching of circuit from one logic 1 to logic 0 and vice-versa.

Q9. What are the methods used for optimization of switchig activity?

Ans:

The various methods that are used to optimize or reduce the switching activity in a CMOS circuit are,

1. Algorithmic optimization
2. Glitch reduction
3. Gated clock signals

Q10. What are the various techniques used to reduce the switched capacitance or parasitic capacitance in a circuit?

Model Paper-3, Q1(g)

Ans:

Switched capacitance or parasitic capacitance is one of the main factors causing power dissipation in a circuit. This capacitance can be optimized using various techniques at various levels. They are,

1. **System Level :** The technique used at system level to optimimize the parasitic capacitance is limiting the usage of shared resources among various devices.
2. **Circuit Level :** The technique used at circuit level to optimimize the parasitic capacitance is changing the logic style to design a circuit or network.
3. **Mask level:** The technique used in mask level or physical design level to optimimize the parasitic capacitance is minimizing the dimensions of the transistors in the circuit.

PART-B ESSAY QUESTIONS WITH SOLUTIONS

2.1 Low-Power Design Approaches

2.1.1 Low-Power Design Through Voltage Scaling – VTCMOS Circuits, MTCMOS Circuits

Q11. Explain the influence of voltage scaling on power dissipation and propagation delay.

Ans:

Model Paper-

The average dynamic switching power dissipation is proportional to the square root of the power supply voltage, expressed as,

$$P_{avg} = \frac{1}{T} C_{load} V_{DD}^2 \text{ (or) } P_{avg} = C_{load} V_{DD}^2 f_{CLK}$$

So, reducing V_{DD} will drastically reduce the power consumption.

- The influence of power supply voltage scaling on power consumption and delay can be observed from the follow propagation delay expressions of CMOS inverter circuit,

$$\tau_{PHL} = \frac{C_{load}}{k_n (V_{DD} - V_{T,n})} \left[\frac{2V_{T,n}}{V_{DD} - V_{T,n}} + \ln \left(\frac{4(V_{DD} - V_{T,n})}{V_{DD}} - 1 \right) \right]$$

$$\tau_{PLH} = \frac{C_{load}}{k_n (V_{DD} - V_{T,n})} \left[\frac{2V_{T,p}}{V_{DD} - |V_{T,p}|} + \ln \left(\frac{4(V_{DD} - |V_{T,p}|)}{V_{DD}} - 1 \right) \right]$$

- From the equations, it is observed that decrease in power supply voltage reduces the dynamic (switching) power dissipati or but increases the propagation delay.

Figure (1) represents the influence of voltage scaling on power dissipation and propagation delay,

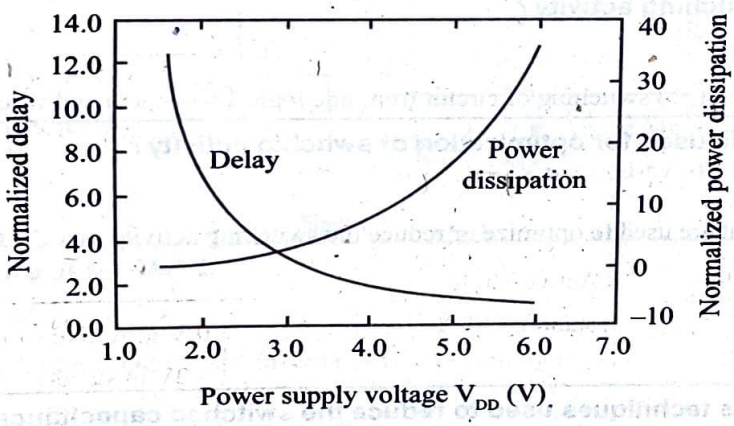


Figure (1): Normalized Variation of Propagation Delay and Average Switching Power Dissipation as a Function of V_{DD}

- Since the speed of the circuit also depends on power supply, it also affects the power dissipation.
- On observing equation (1), it can be assumed that at a constant switching frequency or transition factor, the power con sumption is reduced as quadratic function when V_{DD} is reduced.
- If the CMOS circuit is operated continuously at its highest frequency allowable by its propagation delay, then switchin frequency will decrease as propagation delay increases on power supply reduction.

Figure (2) illustrates the propagation delay for various values of V_T and V_{DD}.

- If the threshold voltage, V_T is scaled down, then it can compensate for the propagation delay when supply voltage is re duced.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

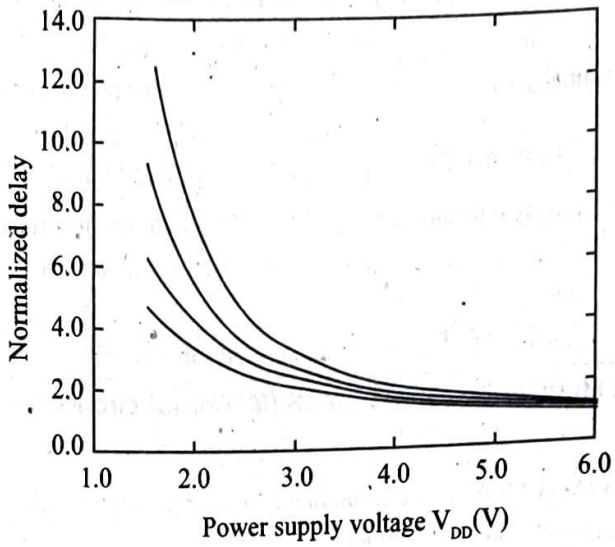


Figure (2)

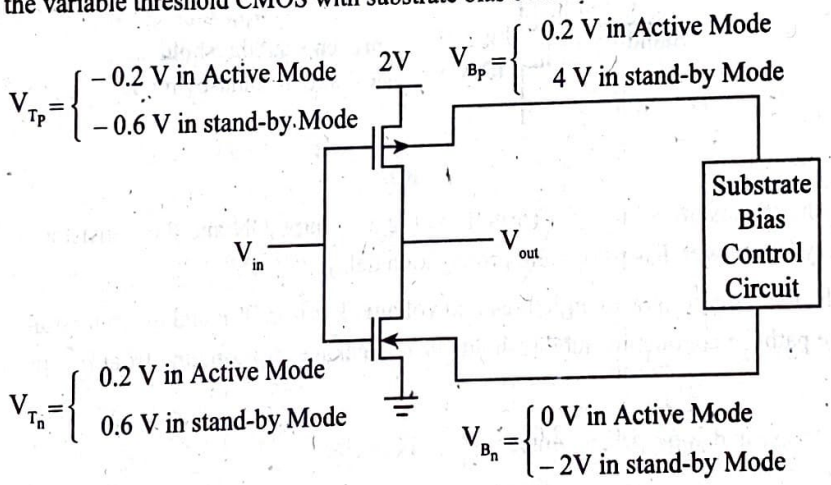
- From figure (2), the scaling of threshold value, V_T from 0.8 to 0.2 reduces the delay value by factor 2 at $V_{DD} = 2$.
- Despite of reduction in delay, scaling of threshold voltage leads to small noise margins and subthreshold current leakages in CMOS circuits.

Q12. Explain briefly about Variable Threshold CMOS (VTCMOS) circuits.

Model Paper-2, Q4

Ans:

- Variable Threshold CMOS (VTCMOS) circuits are designed to overcome the drawbacks of low threshold voltage valued circuits such as subthreshold leakage and power dissipation.
- Figure illustrates the variable threshold CMOS with substrate bias control circuit.


V_{T_P} = – 0.2 V in Active Mode / – 0.6 V in stand-by Mode
2V
V_{B_P} = 0.2 V in Active Mode / 4 V in stand-by Mode
Substrate Bias Control Circuit
V_{in}
V_{out}
V_{T_n} = 0.2 V in Active Mode / 0.6 V in stand-by Mode
V_{B_n} = 0 V in Active Mode / – 2V in stand-by Mode

Figure

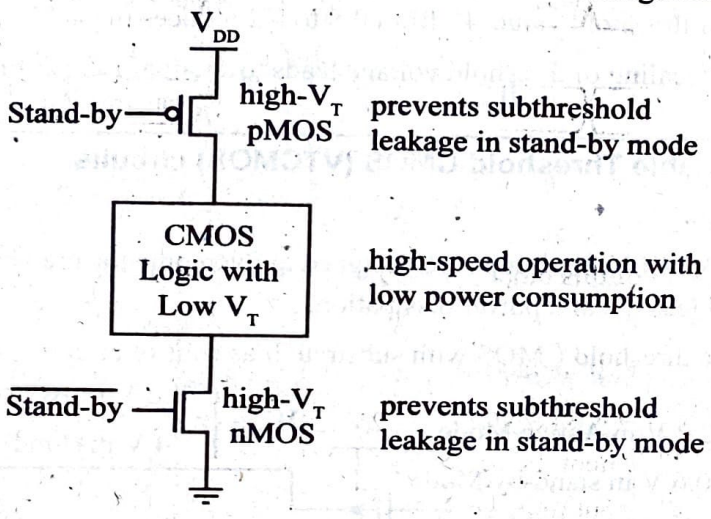
- In a conventional CMOS circuit, all substrate terminals of nMOS transistors are connected with ground, whereas that of pMOS transistors with power supply voltage, V_{DD}. For a MOS transistor, the threshold voltage depends on substrate voltage, V_{SB}.
- In VTCMOS method, the transistors are designed with very less threshold voltages and substrate voltage of these transistors are produced using a variable substrate bias control circuit.
- The self-substrate bias circuit controls the body bias to achieve different threshold voltages.
- If this inverter is operating in active mode, then V_{SB} of nMOS transistor is $V_{Bn} = 0$ and of pMOS transistor is $V_{Bp} = V_{DD}$.
- In active mode, a nearly zero bias is applied. In this mode, the transistors are not affected by back-gate effect. It operates at low power supply and threshold voltages with high switching speed and low power dissipation.
- If the inverter is operating in stand-by mode, the substrate bias control circuit provides low V_{SB} for nMOS and high V_{SB} for pMOS.

- In standby mode, a deeper reverse body bias is applied to increase threshold voltage and cut-off leakage current.
- In this mode, transistors are affected by back-gate effect resulting in high threshold voltages, V_{Tn} and V_{Tp}. As the subthresh leakage current drops exponentially with increasing threshold voltages, the power dissipation decreases consequently.
- Using VTCMOS approach, the threshold voltages can be regulated automatically in order to reduce leakage currents a compensate for the process variations. This technique is referred as Self-Adjusting Threshold Voltage Scheme (SATS
- Though, the VTCMOS technique is efficient in designing CMOS circuits with reduced subthreshold current leakage a adjustable threshold values in the range of V_{DD} - low V_T, it requires a twin-well or triple-well CMOS process to indepe dently control voltage in different parts of chip.
- It also requires a substrate bias generator to generate the rquired substrate bias.

Q13. Explain in detail about Multiple Threshold CMOS (MTCMOS) circuits with necessary circuit diagram

Ans: Model Paper-3,

- Multiple Threshold CMOS (MTCMOS) is the technique used to decrease the subthreshold current leakages and pow dissipation in the circuit operating in stand-by mode.
- This technique uses two transistors with two or more different threshold voltages as shown in figure.



Figure

- In active mode, the transistors with high threshold voltage V_T turns ON and the transistors with low threshold voltag operates with very less power dissipation and propagation delay.
- In stand-by mode, the transistors with high threshold voltage, V_T turns OFF and the transistors with low threshold voltag turns ON and the path for conducting subthreshold current leakages (originating from low threshold circuits) are cut off

Advantage

- It is very simple to use and apply when compared to VTCMOS.

Disadvantages

- In this technique, all the stand-by transistors connected in series increases the total circuit area.
- This method also introduces parasitic capacitance and delay in the circuit.

2.1.2 Architectural Level Approach – Pipelining and Parallel Processing Approaches

Points to Remeber
The pipelining techniques and parallel processing are two of the most common approaches to compensate for performance loss caused by a lower supply voltage with a constant threshold voltage.

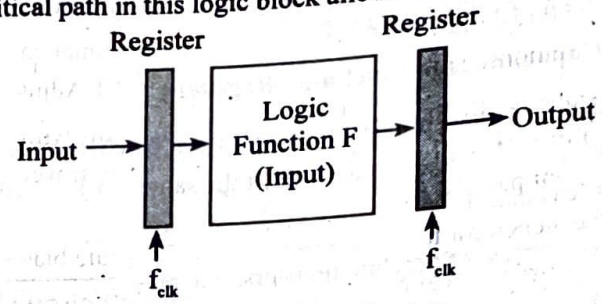
Q14. Explain the pipeling approach.

Ans: Model Paper-1, Q5

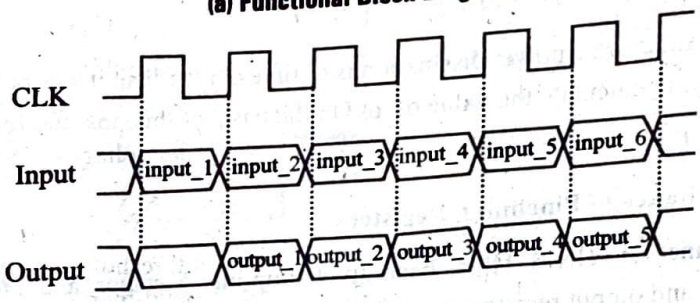
The pipelining approach is a method of applying a data stream in a regular manner to a processing circuit in order to produce results in a sequential order. The supply voltage can be lowered with the pipelining approach, which saves power.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Figure (1) shows the single functional block that implements the logic function F(INFUT) on the input vector, INPUT and its timing diagram. Each of the input and output vectors is sampled through a register array that is driven by a clock signal called CLK. At a voltage of V_{DD}, the critical path in this logic block allows to sample at a maximum sampling rate of f_{clk}.



(a) Functional Block Diagram



(b) Timing Diagram

Figure (1)

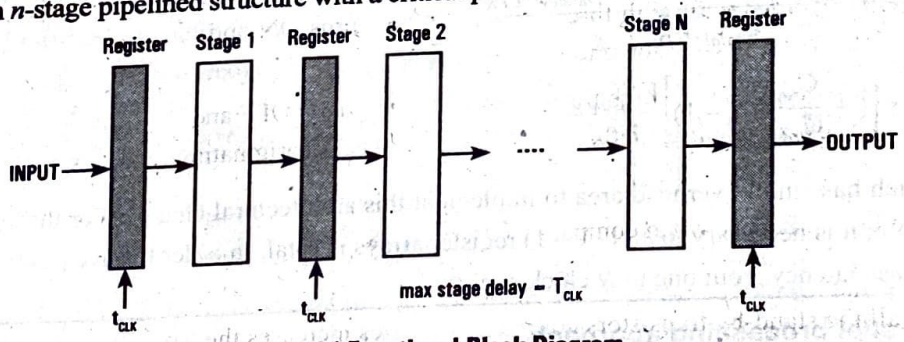
The dynamic power consumption of this block is given by, ... (1)

$$P_{reference} = C_{total} \cdot V_{DD}^2 \cdot f_{CLK}$$

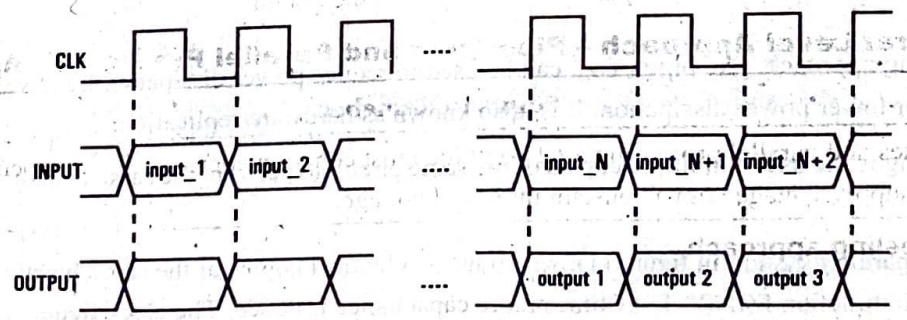
Where, C_{total} is the total capacitance. It includes,

- The capacitance switched in the input register array
- The capacitance switched to implement the logic function
- The capacitance switched in the output register array.

Now, Consider an n-stage pipelined structure with a critical path that is only $1/n$ of its original length as shown in figure (2).



(a) Functional Block Diagram



(b) Timing Diagram

Figure (2)

Here, two cases arise.

1. Without considering the capacitance of pipelining registers
2. Considering the capacitance of pipelining registers.

1. Without Considering the Capacitance of Pipelining Registers

During a single clock period, the amount of capacitance that can be charged or discharged is reduced to C_{total}/n, wi considering the capacitance of pipelining registers. As a result, if the same clock speed is preserved, the voltage can be re to αV where α is a positive number less than 1.

Without considering the capacitances of pipeline registers, the power dissipation of the pipelined structure is given

$$P_{pipeline} = n \times \frac{C_{total}}{n} \times f \times \alpha^2 V^2 = \alpha^2 P_{seq}$$

As a result, the pipelined structure's power dissipation is α^2 times lower than it was before. The propagation delay of structures is assumed to be equal, to calculate the value of α. On the basis of this, that the following must be valid.

$$n(\alpha V - V_T)^2 = \alpha(V - V_T)^2$$

2. Considering the Capacitance of Pipelining Registers

In figure (2), the logic function F(INPUT) has been divided up into N stages, and a total of $(N-1)$ register arrays h been added to the original input and output registers to make the pipeline. All registers are clocked at the same rate, f_{clk}.

With a lower supply voltage and the same functional throughput as the single-stage structure, the dynamic po consumption of the N-stage pipelined structure can be approximated by,

$$P_{pipeline} = [C_{total} + (N-1)C_{reg}].V_{DD.new}^2.f_{CLK}$$

Where ,

C_g – The capacitance switched by each pipeline register.

Then, the power reduction factor achieved in a N-stage pipeline structure is,

$$\frac{P_{pipeline}}{P_{reference}} = \frac{[C_{total} + (N-1)C_{reg}].V_{DD.new}^2.f_{CLK}}{C_{total}.V_{DD}^2.f_{CLK}}$$

$$= \left[1 + \frac{C_{reg}}{C_{total}}(N-1)\right]\frac{V_{DD.new}^2}{V_{DD}^2}$$

Pipelining approach has small overhead area to implement this architectural change. For the single-stage structure to transformed into a pipeline, it is necessary to add $(N-1)$ register arrays in total. In order to save power, this technique trades area . However, it increases latency from one to N clock cycles.

Q15. Explain the parallel processing approach.

Ans: Model Paper-I

- Parallel processing approach, like pipelining, can be used to reduce power dissipation by lowering the supply voltage trades-off area for lower power dissipation. It is also known as hardware replication.
- Parallel processing is the use of multiple copies of the same circuit to perform the same computation on multiple data s simultaneously.
- Consider the N-parallel system in figure (1), which has N identical copies of the same function units each of which plements the logic function F(INPUT). N times more capacitance is added. The clock frequency must be decreased by times in order to maintain the same level of performance. Since the N capacitors (NC_{total}) have more time to charge discharge, the supply voltage can be decreased by a factor α.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

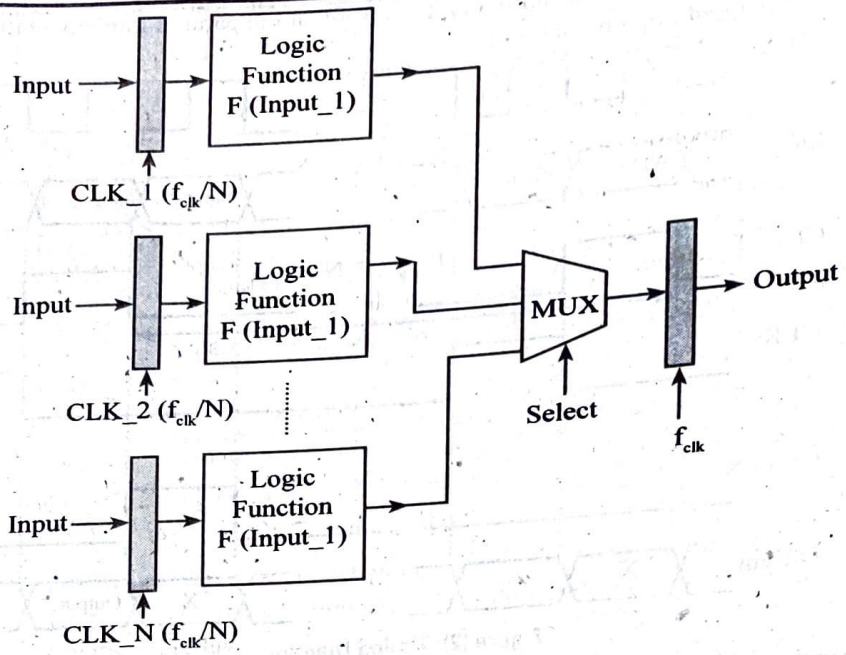


Figure (1)

- Assume that the input vectors arrive at the same rate. All the registers of the N processing blocks receive the input vectors. In order to load each register per (NT_{clk}) clock cycles, a gated clock signal is used. Since the clock signals to each input register are biased by the T_{CLK}, the N consecutive input vectors are loaded into a new register for each N consecutive inputs.
- As a result, the computation time for each input vector is increased by a factor of N since each input register is clocked at a lower frequency (f_{CLK} /N). The power supply voltage can be lowered until critical path delay equals the new clock period of NT_{clk}.
- The total dynamic power dissipation of the parallel structure (neglecting the dissipation of the multiplexer) is equal to the sum of the power dissipated by the input registers and the logic blocks operating at a clock frequency of (f_{CLK}/N) and the output register operating at a clock frequency of f_{CLK}. It is given by,

$$P_{parallel} = N.C_{total}.V_{DD.new}^2.\frac{f_{CLK}}{N} + C_{reg}.V_{DD.new}^2.f_{CLK}$$

$$P_{parallel} = \left(1 + \frac{C_{reg}}{C_{total}}\right).C_{total}.V_{DD.new}^2.f_{CLK} \quad \text{... (1)}$$

Ignoring the additional overhead, associated input routing capacitance, the output routing capacitance and the capacitance of the output multiplexor structure, the amount of power reduction achievable in a N-block parallel implementation is,

$$\frac{P_{parallel}}{P_{reference}} = \frac{V_{DD.new}^2}{V_{DD}^2}.\left(1 + \frac{C_{reg}}{C_{total}}\right) \quad \text{... (2)}$$

The lower bound of switching power reduction realizable with architecture-driven voltage scaling assuming zero threshold voltage is given by,

$$\frac{P_{parallel}}{P_{reference}} \geq \frac{1}{N^2} \quad \text{... (3)}$$

Without considering the capacitances of parallel-processing registers, the power dissipation of the m-parallel system can be expressed as,

$$P_{parallel} = (m_{Ctotal}) \times \frac{f}{m} \times \alpha^2 V^2 = \alpha^2 P_{seq} \quad \text{... (4)}$$

Figure (2) illustrates the timing diagram of figure (1). It shows that the parallel implementation has a latency of N cycles, just like in the *N*-stage pipelined implementation.

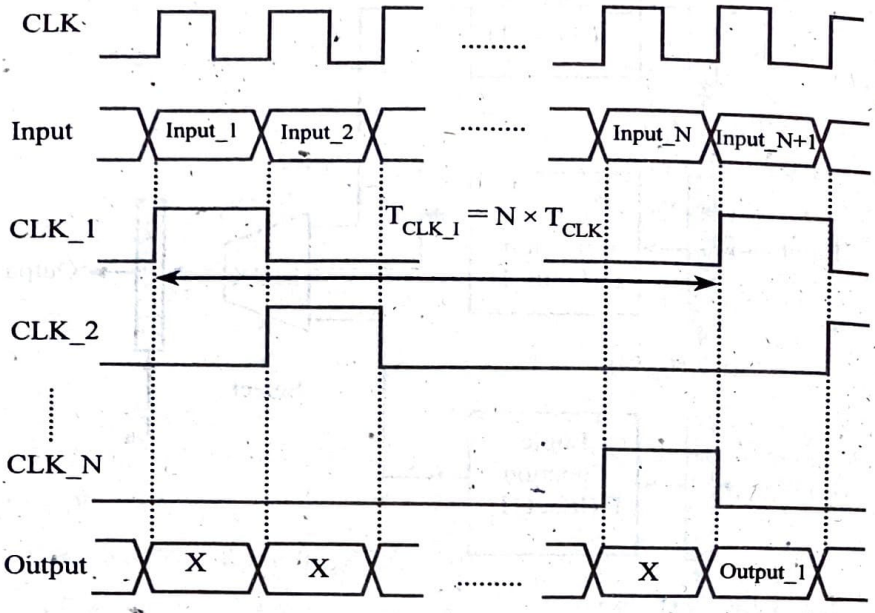


Figure (2): Timing Diagram

2.2 Switched Capacitance Minimization Approaches – System Level Measures, Circuit Level Measures and Mask Level Measures

Q16. Explain briefly about the estimation of switching activity.

Ans:

The switching activity refers to switching of circuit from one logic 1 to logic 0 and vice-versa.

- In switching power dissipation, the power consumption or dissipation of CMOS circuit depends on switching activity factor which in turn depends on Boolean function of the logic gate.
- To illustrate the concept of switching activity consider a 2 input CMOS NOR2 gate.
- If P_0 and P_1 are signal probabilities, where P_0 represents the probability of occurrence of logic 0 and P_1 represents the probability of occurrence of logic 1 at the output, then the probability of occurrence of power consuming transition at the output is the product of these two probabilities i.e., P_0 and P_1.
- It hs four possible state (00, 01, 10, 11) for inputs x and y. From the truth table.

$P_0 = 3/4$ and $P_1 = 1/4$

The probability of occurrence of power consuming transition at the output of NOR gate is,

$$P_{0\to1} = P_0 . P_1 = \frac{3}{4} \cdot \frac{1}{4} = \frac{3}{16}$$

Figure (1) represents the state transition diagram of a CMOS NOR 2 gate.

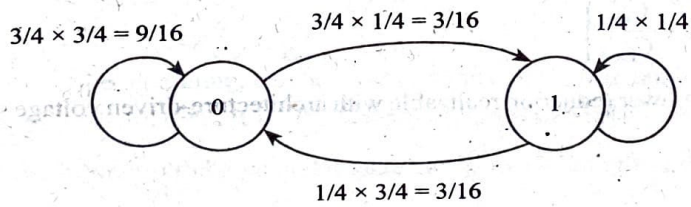


Figure (1)

- Generally, the probability of occurrence of power consuming transition at the output of an n-input CMOS logic is a function of n_0 and is given as,

$$P_{0-1} = P_0 . P_1 = \left(\frac{n_o}{2^n}\right) . \left(\frac{2^n - n_o}{2^n}\right)$$

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Figure (2) illustrates the probability of power consuming output transition as a function of number of inputs for XOR, XNOR, NAND and NOR gates.

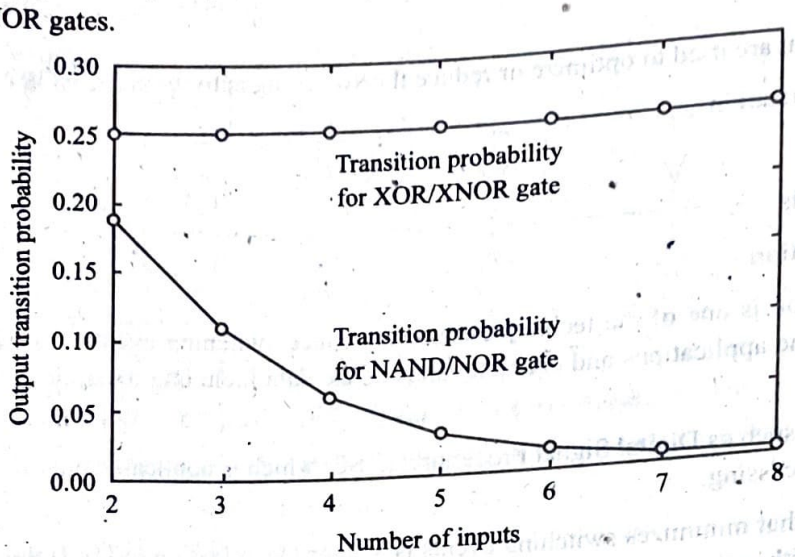


Figure (2)

If a NOR gate has two inputs which are independent and not uniformly distributed, then it is not possible to estimate the occurrence of logic 0 and 1 with same probabilities. Then, the probability of power consuming output transition is the function of input probability distributions.

Where,

The probability of occurrence of logic 1 at input $A = P_{1,A}$

The probability of occurrence of logic 1 at input $B = P_{1,B}$

Then,

The probability of occurrence of logic 1 at output is, ... (2)

$$P_1 = (1 - P_{1,A})\,(1 - P_{1,B})$$

From equation (2), the probability of occurrence of power consuming transition at output is given as,

$$P_{o\to 1} = P_0 . P_1 = (1 - P_1).\, P_1 \quad \text{... (3)}$$

$$P_{o\to 1} = (1 - (1 - P_{1,A}).\,(1 - P_{1,B}))\,((1 - P_{1,A}) \,.\, (1 - P_{1,B}))$$

Therefore, equation (3) represents the probability of power consuming output transition as a function of input probability distributions and this distribution is shown in figure (3).

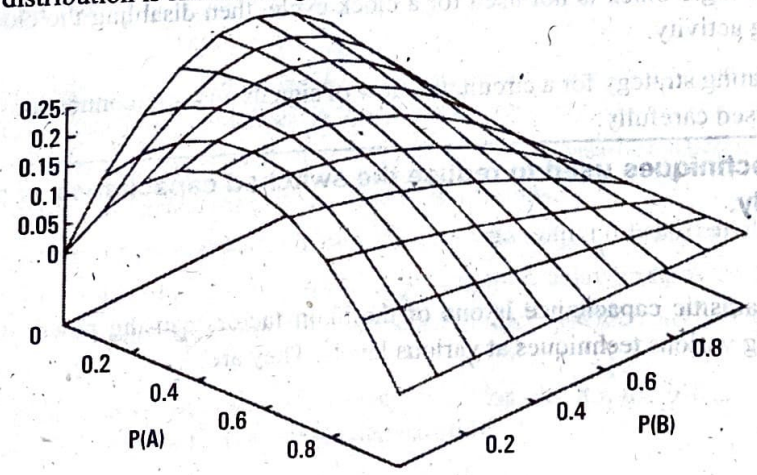


Figure (3)

Therefore, from equation (3) it can be observed that, it is little complex to evaluate the switching activity in multilevel circuits involving sequential components, feedback loops,etc. So, a designer prefers CAD tools to estimate the switching of any network.

Q17. What are the methods used for optimization of switching activity? Explain them briefly.

Ans:

The various methods that are used to optimize or reduce the switching activity in a CMOS circuit are,

1. Algorithmic optimization
2. Glitch reduction
3. Gated clock signals

1. Algorithmic Optimization

- Algorithmic optimization is one of the techniques used to reduce switching events in CMOS circuits. This techn mainly depends upon the applications and characteristics of the data including dynamic range, correlation and statis of data transmission.
- There are few methods such as Digital Signal Processing (DSP) which is applicable only for certain applications and for general purpose processing.
- One of the algorithms that minimizes switching events is Vector Quantization (VQ). If this algorithm employs a diff ential tree search algorithm instead of full search algorithm, then it can also optimize the number of memory acces multiplications and additions by a factor of 30.
- At system level, switching activity can be reduced by changing the data representation. Switching events can be optim greatly while representing a signed data, if sign-magnitude representation is used instead of two's complement represen tion.

2. Glitch Reduction

- Glitches are the fast signal transitions caused due to the carry of propagation delays from one block to the other blo These glitches results in critical races or dynamic hazards.
- If input signals of a gate does not transit from ground to V_{DD} at the same time, it results in dynamic hazards. If the in signals transit partially, it results in partial hazards. If the input signals transit from ground to V_{DD} at the same time, th no glitches occur in the output.
- The main reason behind the occurrence of hazards is improper balance in the lengths of paths between gates in a netwo This leads to the improper timing or propagation delays in transmitting signals between logic blocks.

3. Gated Clock Signals

- Using of gated clock signals or conditional clock signals is another way of reducing switching activity in the network
- In a network, if a particular logic block is not used for a clock cycle, then disabling the clock signal for that particu block reduces the switching activity.
- In order to design a clock gating strategy for a circuit, the flow of signals and interconnections among the blocks and th operations have to be analysed carefully.

Q18. What are the various techniques used to reduce the switched capacitance or parasitic capacitance a circuit? Explain briefly.

Ans:

Model Paper-3, I

Switched capacitance or parasitic capacitance is one of the main factors causing power dissipation in a circuit. Th capacitance can be optimized using various techniques at various levels. They are,

1. System level
2. Circuit level and
3. Mask level

1. System Level

- The technique used at system level to optimimize the parasitic capacitance is limiting the usage of shared resour among various devices.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Example

Figure (1) illustrates various modules using a single bus to access CPU resources. It is a global bus structure for data transmission between multiple modules.

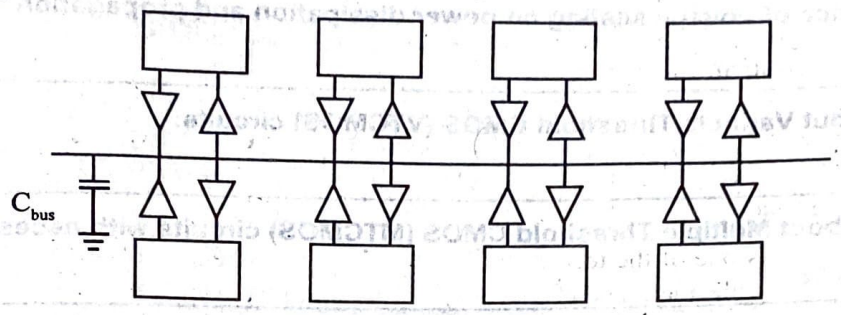


Figure (1): Single Global Bus Structure for Connecting a Large Number of Modules on a Chip

Since a single bus is shared among various modules, a large bus capacitance occurs due to following reasons,

(i) Usage of same bus by large number of drivers and receivers.

(ii) Parasitic capacitance of the long bus line.

- In order to access the bus, high amount of power is consumed to drive this large capacitance of the bus.
- The parasitic capacitance or switched capacitance can be optimized by dividing the single global star bus into small multiple buses as shown in figure (2).

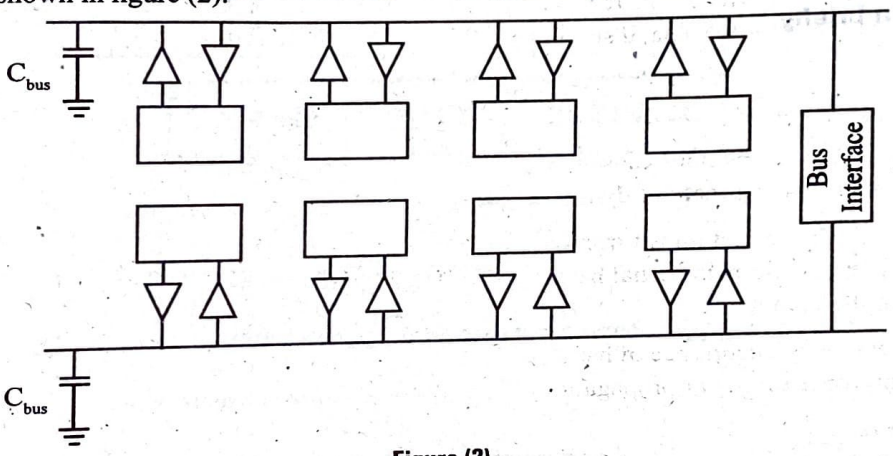


Figure (2)

- These multiple small buses regulate the data transmission among various modules and also reduce the switching capacitance while accessing the bus.

2. **Circuit Level**

- The technique used at circuit level to optimimize the parasitic capacitance is changing the logic style to design a circuit or network. One of the main reasons for the increase in load capacitance in the circuit is the usage of large number of transistors.
- One method to minimize this load capacitance is to use transfer gates pass-transistor logic in the place of CMOS gates.
- Even though it reduces the load capacitance, it has several limitations,

(i) Pass-gates require inverters at the output in order to generate output driving capabilities in acceptable range.

(ii) Usage of inverters results in increase of total area, propagation delay and power dissipation.

(iii) It uses dual-rail logic to generate complementary control signals which in turn requires two more complementary nMOS-pass transistors.

- All these disadvantages overcome the advantage of pass-tranistor. So, a trade off must be maintained among required factors and the required pass-transistor has to be chosen carefully.

3. **Mask Level**

- The technique used in mask level or physical design level tooptimimize the parasitic capacitance is minimizing the dimensions of the transistors in the circuit.
- Though small-sized transistors in a logic circuit reduce the power consumption, it affects the dynamic performance of the network. Therefore, a trade-off must be maintained between power consumption and dynamic performance in mask level.

IMPORTANT QUESTIONS

Q1. Explain the influence of voltage scaling on power dissipation and propagation delay.

Important Questi

Ans: Refer Q11.

Q2. Explain briefly about Variable Threshold CMOS (VTCMOS) circuits.

Important Questi

Ans: Refer Q12.

Q3. Explain in detail about Multiple Threshold CMOS (MTCMOS) circuits with necessary circuit diagram.

Important Questi

Ans: Refer Q13.

Q4. Explain the pipeling approach.

Important Questi

Ans: Refer Q14.

Q5. Explain the parallel processing approach.

Important Questi

Ans: Refer Q15.

Q6. What are the various techniques used to reduce the switched capacitance or parasitic capacitance in a circuit? Explain briefly.

Important Questi

Ans: Refer Q18.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

UNIT 3

Low-Voltage Low-Power Adders

Look for the Logo on the Title Cover Before You Buy
SIA GROUP
SIA GROUP

Syllabus

LOW-VOLTAGE LOW-POWER ADDERS

Introduction, Standard Adder Cells, CMOS Adder's Architectures – Ripple Carry Adders, Carry Look- Ahead Adders, Carry Select Adders, Carry Save Adders, Low-Voltage Low-Power Design Techniques –Trends of Technology and Power Supply Voltage, Low-Voltage Low-Power Logic Styles.

Learning Objectives

- Need of Low-voltage Low-power Adders
- Standard cell, schematic configurations of Full adders
- Basic operation of ripple carry adder and its performance evaluation, glitching problem
- Basic theory and operation of carry look-ahead adder
- Operation of 16-bit Carry Look-Ahead adder, ELM adder and carry save adder (CSA)
- Operation of carry select adder (CSL) and its performance evaluation relative to RCA
- Implementation of XOR/XNOR Gate using different logic styles

Introduction

It is extremely crucial to conduct thorough research on adders before designing a circuit. The fundamental arithmetic operations would be incomplete without addition, which is a necessary and fundamental function. General-purpose systems and application-specific processors alike rely on it heavily, as it appears in many VLSI design paradigms. Since each bit must transmit a carry signal to all higher bit positions, this is a very important operation that must be done correctly. Finally, the chapter ends with multiple-valued logic-based adders. High-speed addition is possible due to this new signed-digit number system and the residue number system, which were both carefully chosen arithmetic representations of numbers.

PART-A SHORT QUESTIONS WITH SOLUTIONS

Q1. What is standard adder cell?

Ans:

Model Paper-

Standard adder cells are the basic building blocks utilized in designing and fabricating many sorts of adder archite We have two types of standard adder cells,

1. Hall adder
2. Full adder

This means the half adder and full adder are used as building blocks in the design of complex adder structures.

Q2. What is the need of Low-voltage Low-power Adders? Explain.

Ans:

The fundamental mathematical calculations cannot be performed without doing addition, addition is an indispen operation. Also, subtraction, multiplication, division, and address computation all depend on the addition operation. He addition is regarded as an indispensable component of the arithmetic unit. Hence, Low-voltage Low-power Adders are develo

Q3. How carry propagation time can be minimized for ripple carry adder.

Ans:

Model Paper-1, Q

The carry propagation time of ripple carry adder can be minimized by using various designs of improved full ad standard cell. The use of fastest full adder standard cell in the construction of ripple carry adder minimizes the carry propaga time. Designing ripple carry adder using the 10-T FA model instead of TFA model exhibits 44% more speed. Also, 10-T model performs better than the TFA model over the entire power supply range.

Q4. What is an ELM adder.

Ans:

Model Paper-2, Q

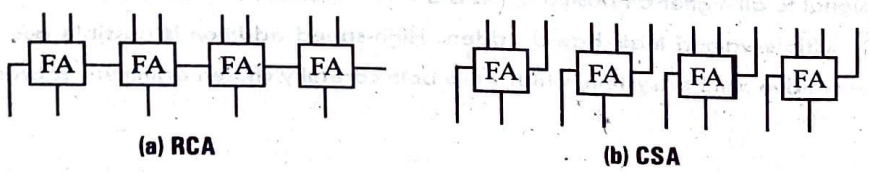
An ELM adder is a variation of the basic Carry Look-Ahead adder algorithm. The algorithm of ELM addition encompass a binary tree of basic processors that run in O(log n) time. The ELM algorithm was developed using the idea of carry propag and generate. The ELM adder directly calculates the sum bits in parallel, which allows it to lower the number of inter-c connections.

Q5. What is carry save adder (CSA) and how is it obtained.

Ans:

Model Paper-2, Q

Carry Save Adder (CSA) is in fact a Ripple Carry Adder (RCA) with its carries saved instead of propagated. In oth words, Carry Save Adder is obtained by slight modification of the Ripple Carry Adder, that is, the carrier is saved within t full adder instead of propagating it to next full adder block. Figure (1) illustrates the block diagrams of both RCA and CSA f comparison. The Carry Save Adder performs concurrent addition of multiple inputs.



(a) RCA (b) CSA

Figure (1): Block Diagram

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

6. Explain static and dynamic logic styles. Model Paper-3, Q1(c)

ns:

There are two types of CMOS logic circuits: static and dynamic. The output of static logic families is calculated when e input changes, whereas the output of dynamic logic gates is calculated only once per every clock cycle. Dynamic gates are ifferent from static gates because they are clocked and perform in the precharge and evaluation phases.

7. Distinguish between Transmission Function Full Adder and 10 Transistor Full adder. Model Paper-3, Q1(h)

ns:

The differences between transmission function full adder and 10 Transistor Full adder are tabulated below.

	Transmission Function Full Adder (TFA)		10-Transistor Full adder (10-T FA)
1.	TFA encompasses 16 transistors	1.	10-T FA encompasses 10 transistors
2.	When the voltage of the power supply is set to 3.3 V, the critical path delay of the TFA measures 0.12 ns.	2.	When the voltage of the power supply is set to 3.3 V, the critical path delay of the 10-transistor FA measures 0.086 ns
3.	At a voltage of 3.3 V and a clock frequency of 1 GHz, the TFA dissipates around 170 μW.	3.	At a voltage of 3.3 V and a clock frequency of 1 GHz, the 10-transistor FA consumes an average of 81 μW of power.

Q8. Why carry look-ahead adder is developed?

Ans:

In the case of a ripple carry adder, the size and delay associated with the logic circuit increase linearly with input operands. That is, as input increases, the size and delay of the adder increases. In order to overcome this drawback, the carry look-ahead adder has been developed.

Q9. What are the different logic styles used in the implementation of XOR/XNOR gate.

Ans:

Five different logic styles are explicitly used for implementation of XOR/XNOR gate. They are,

1. Full static CMOS
2. Complementary Pass-transistor Logic (CPL)
3. Double Pass-transistor Logic (DPL)
4. Dual-rail Domino dynamic logic
5. Single-rail Domino dynamic logic.

PART-B ESSAY QUESTIONS WITH SOLUTIONS

3.1 Introduction, Standard Adder Cells

Q10. What is the need of Low-voltage Low-power Adders? Explain.

Ans:

Model Paper-3, Q

The fundamental mathematical calculations cannot be performed without doing addition, addition is an indispensab operation. Perhaps the general-purpose system and application-specific processors use the addition operation most recurrently a also it is used extensively in many VLSI systems. Furthermore, subtraction, multiplication, division, and address computation depend on the addition operation. Hence, addition is regarded as an indispensable component of the arithmetic unit. Therefo adder is considered as the heart (center) of each and every microprocessor, DSP architecture, or data processing system. Hen Low-voltage Low-power Adders are developed.

Further, addition is quite important and necessary operation as it encompasses a carry ripple step that propagates a car signal from more or less every bit to its higher bit position. This causes a significant circuit delay. As a result, the adder in t crucial delay path effectively controls the total speed of the system. The main driving force behind the development of lo voltage low-power adders is portable devices such as smart phones, tablets, flash memory devices, handheld mobile telephone personal digital assistants (PDAs), and laptop computers.

Q11. What is standard adder cell? Explain about standard half adder cell.

Ans:

Model Paper-1, Q

Standard adder cells are the basic building blocks utilized in designing and fabricating many sorts of adder architectur We have two types of standard adder cells

1. Hall adder
2. Full adder

This means the half adder and full adder are used as building blocks in the design of complex adder structures.

Standard Half Adder Cell

Adders come in a variety of kinds, but the simplest and most basic of all is the Half Adder. It is composed of two bina operands, each of which receives single bits as input, and as the result of the operation it generates a two-bit binary number. the two-bit binary number, the bit in the LSB position is referred to as "Sum," and the bit in the MSB position is called "Carr Figure (1) depicts the logic gate of the half adder, and Table (1) is the truth table that describes how the half adder gate operat

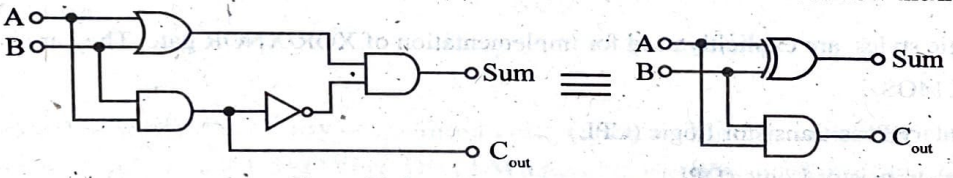


Figure (1): Logic gate of HALF Adder

A	B	Sum	C_{out}
0	0	0	0
0	1	1	0
1	0	1	0
1	1	0	1

Table (1): Truth table of HALF Adder

Based on the truth table, the logic functions for the Sum and Carry can be expressed as follows:

$$Sum = A \oplus B$$

$$C_{out} = A \bullet B$$

Due to the lack of support for carry-in operations, the Half Adder can only be used to add two binary values that eac contains a single bit. In order for the adder to be able to handle the carry-in, it would need to have three inputs rather than jus two. This would allow it to add operands that process more than one bit at a time.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Q12. Explain about Full adder cell and draw the schematic of standard CMOS full adder cell design.

Ans:

In a full adder, two binary numbers are added together with a carry-in. Figure (1) illustrates the structural description of a standard CMOS full adder. It is built with the assistance of two half-adders and an OR gate. The Full Adder takes in a total of three inputs: two for the input numbers A and B, and the other for the carry-in, which is denoted by the letter Cin. As a result of the addition operation, the full adder generates a two-bit binary number. In the resultant two-bit binary number, the bit in the LSB position is referred to as "Sum," and the bit in the MSB position is called "Carry-out," which is denoted as Cout. Table (1) is the truth table that describes how the full adder gate operates.

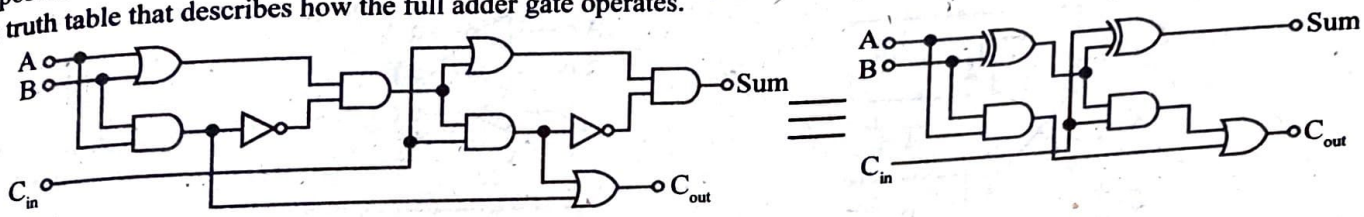


Figure (1): Logic Gate of Full Adder

C_{in}	A	B	Sum	C_{out}
0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	1	1	0	1
1	0	0	1	0
1	0	1	0	1
1	1	0	0	1
1	1	1	1	1

Table (1): Truth Table of Full Adder

Based on the truth table, the logic functions for the Sum and Carry-out (Cout) can be expressed as follows:

$$Sum = A \oplus B \oplus C_{in}$$

$$C_{out} = (A \oplus B) \bullet C_{in} + A \bullet B$$

Figure (2) displays the logic symbol for the FA, and Figure (3) illustrates the implementation of a standard CMOS full adder cell design employing a total of 32 transistors.

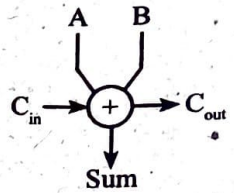


Figure (2): Logic gate of Full Adder

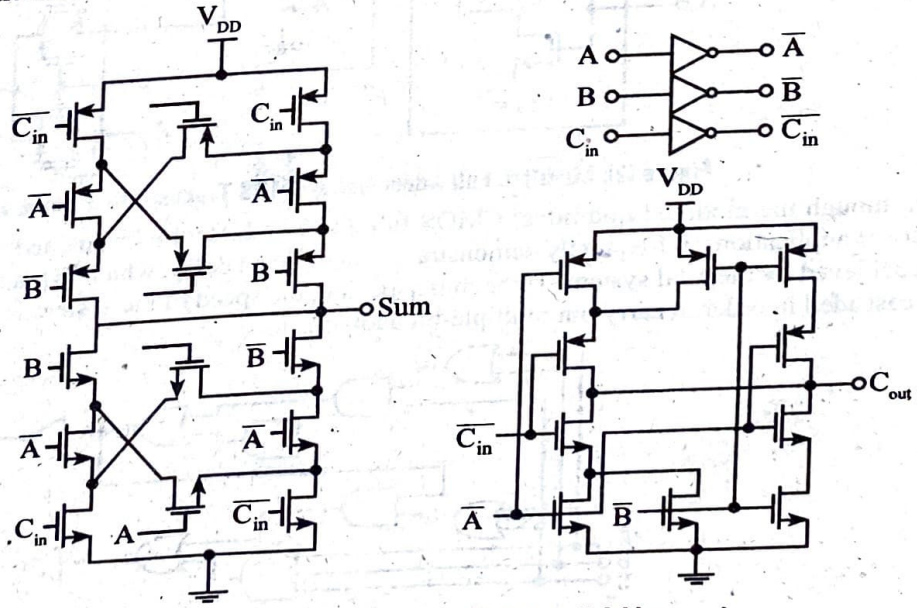


Figure (3): Conventional CMOS Full Adder

Q13. Explain various schematic configurations of Full adders their advantages and disadvantages.

Ans:

Model Paper-2, Q

The realization of a conventional CMOS full adder standard cell design at the transistor level is illustrated in Figure (1). This design makes use of a total of 32 transistors. In order to reduce the space and power consumed by the conventional full adder, it is modified using CMOS transmission gates and inverters as illustrated in figure (2). The modified form contains just 20 transistors. This modification greatly reduces space and power.

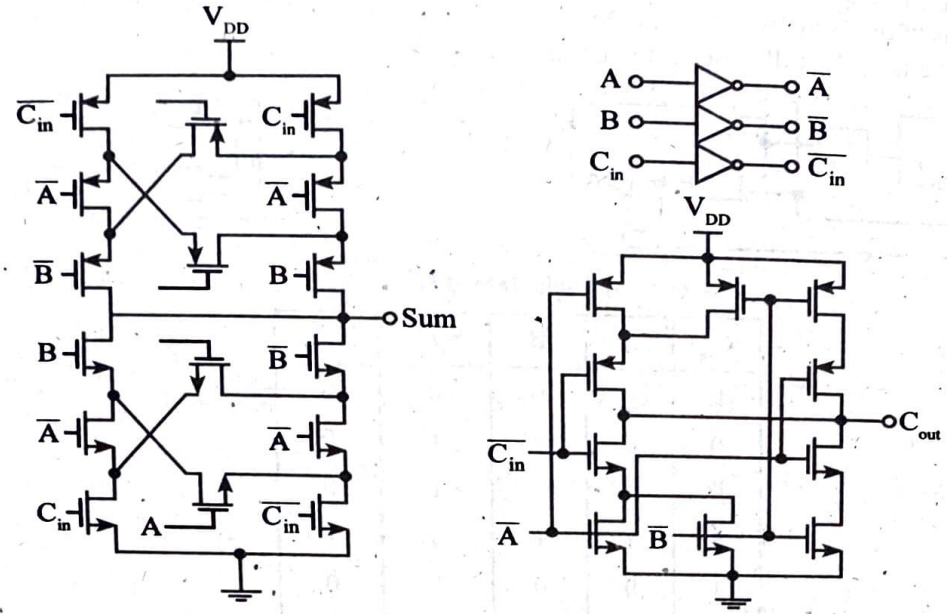


Figure (1): Conventional CMOS Full Adder Circuit

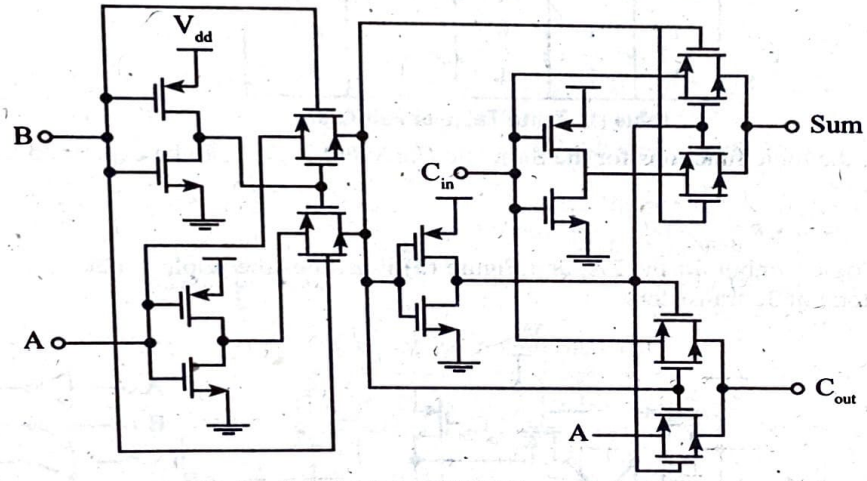


Figure (2): Modified Full Adder Using CMOS Transmission Gates and Inverters

Even though the modified traditional CMOS full adder architecture has gained widespread acceptance and is used in a broad variety of applications, it frequently demonstrates a significant delay, which in practice restricts the amount of performance that can be achieved by the total system. To be more specific, the speed of the system is negatively affected whenever these full adders are cascaded in order to carry out multiple-bit addition.

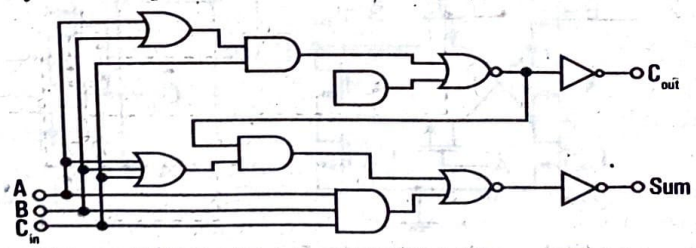


Figure (3): Logic diagram of Full Adder without XOR gates

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

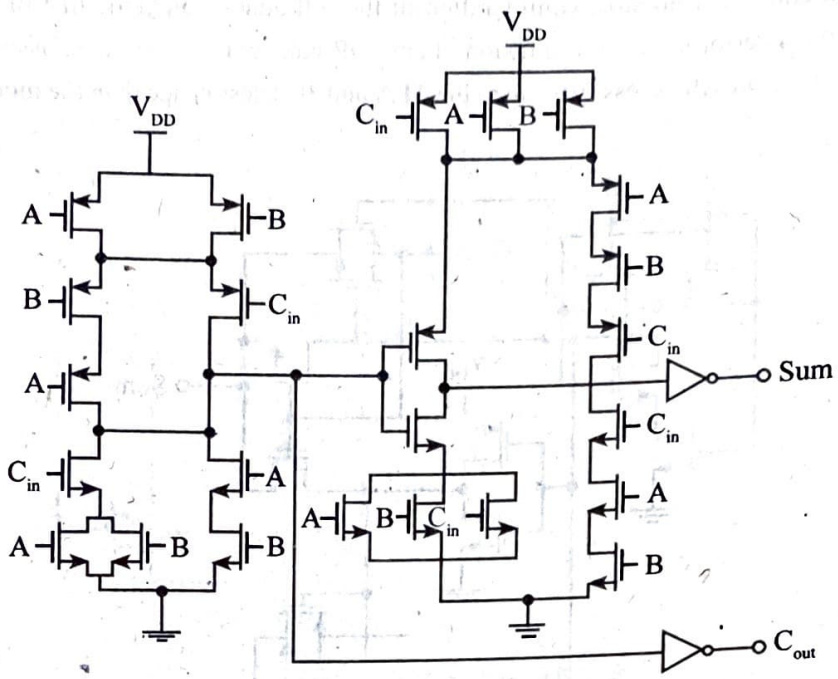


Figure (4): Transistor Diagram of Full Adder without XOR Gates

An alternative to the modified FA is the implementation of the full adder cell using 28 transistors, which doesn't use XOR gates. The logic circuit and transistor diagram of this design are illustrated in figures (3) and (4), respectively. From figure (3) it is clearly evident that this full adder is realized by repeating the C_{out} term in the Sum term as a common sub-expression. The logic functions for the Sum and Carry-out (C_{out}) for this design can be expressed as follows:

$$C_{out} = A \bullet B + C_{in} \bullet (A + B)$$

$$\text{Sum} = A \bullet B \bullet C_{in} + (A + B + C_{in}) \bullet \overline{C_{out}}$$

$$= A \bullet B \bullet C_{in} + (A + B + C_{in}) \bullet \overline{(A \bullet B + C_{in} \bullet (A + B))}$$

Since the adder is the heart (center) of almost all logic systems, the simplification of full-adders is a never-ending process. In this process, a CMOS full adder based transmission function was developed and is illustrated in figure (5). This adder was given the name the Transmission Function Adder (TFA). It encompasses 16 transistors and dissipates less power than conventional CMOS full adders described until now.

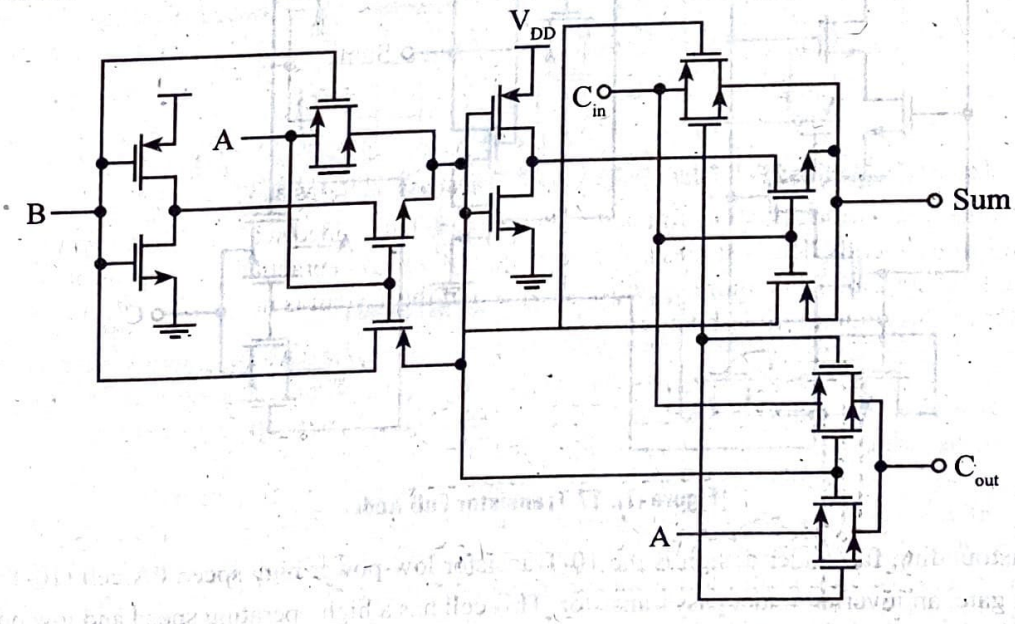


Figure (5): Transmission Function Full Adder (TFA)

Figure (6) illustrates an additional possible configuration of the full adder schematic that offers the advantages of power consumption and fast speed performance. It is a fusion of an XOR gate and several transmission gates all rolled into It is a 14-transistor circuit that consumes 20% less space than the TFA and 30% less space than the modified conventional C full adder.

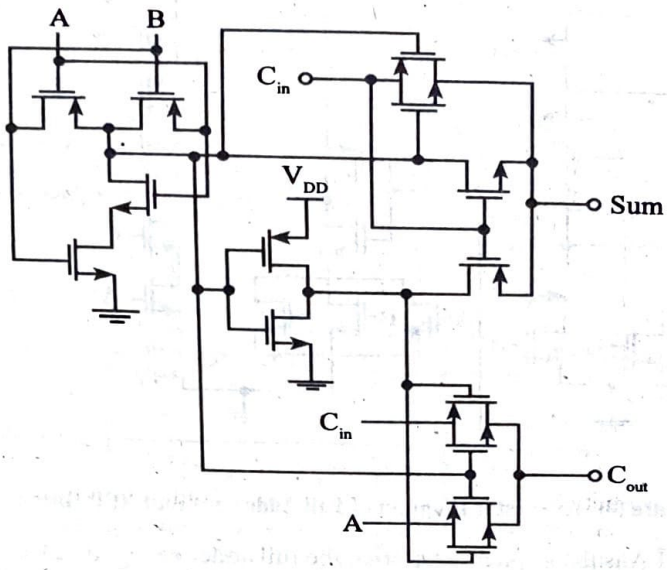


Figure (6): 14 Transistor Full adder

A low-power CMOS full adder cell made up of 17 transistors is now presented. The XOR and XNOR gates, as well as pass transistors, are the foundation of this design. According to comparison tests, it uses 10 - 15% less power than the TFA or t 14-transistor FA. The reason behind this power saving is that there is no short-circuit power dissipation in the 17 transistor ce Also, its dynamic power dissipation is less, compared to the other two cells. Figure (7) illustrates the transistor schematic for t implementation of this design.

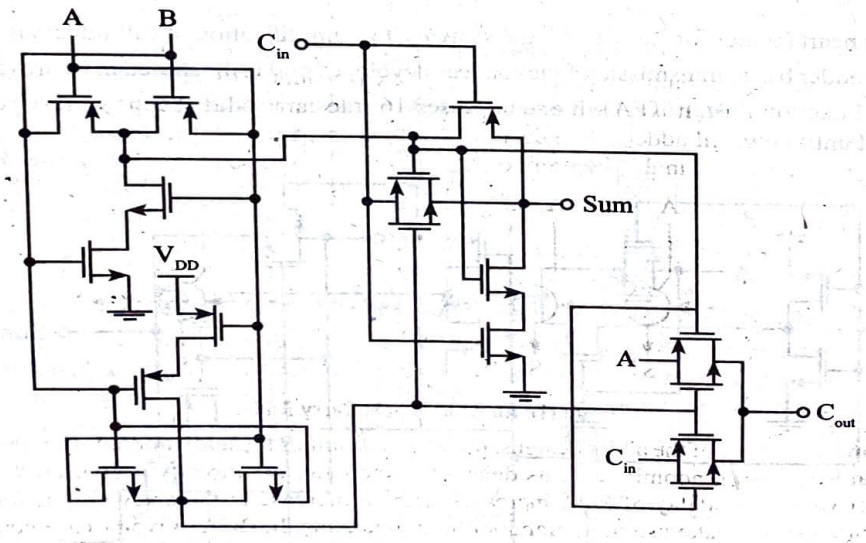


Figure (7): 17 Transistor Full Adder

An even more astounding full adder design is the 10-Transistor low-power high speed FA cell (10-T FA). In its critic path it utilizes an XOR gate, an inverter, and a pass transistor. This cell has a high operating speed and low power consumptio Figure (8) illustrates the schematic diagram of the 10-T FA.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

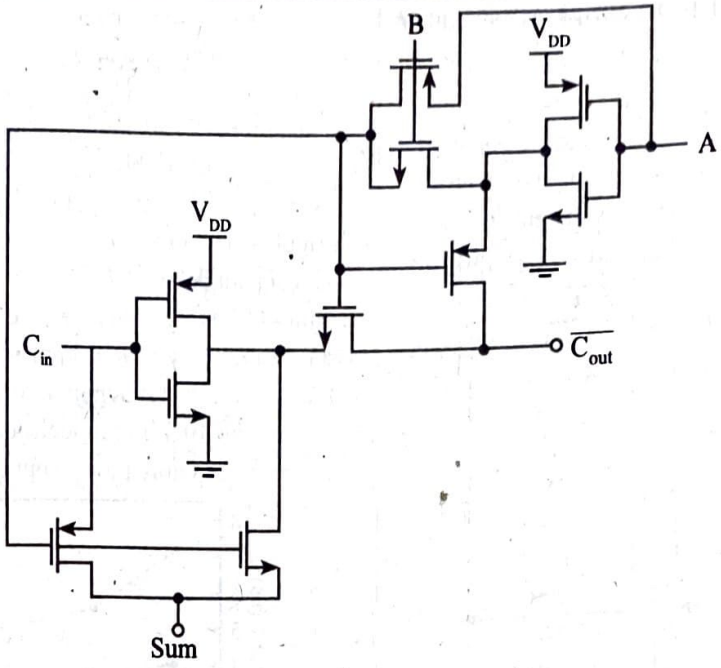


Figure (8): 10 Transistor Full Adder

A comparison has been made between the previously described configuration of the TFA (Transmission Function Full Adder) and this 10-T FA configuration. When the voltage of the power supply is set to 3.3 V, the critical path delay of the 10-transistor FA measures 0.086 ns, whereas the critical path delay of the TFA measures 0.12 ns. Also, with the same supply voltage and a clock frequency of 1 GHz, the 10-transistor FA consumes an average of 81 μW of power, whereas the TFA dissipates around 170 μW. This difference in power dissipation is due to the fact that the TFA has more transistors.

This low-power high-speed full adder cell is used in the design of complex adders like the Ripple Carry Adder (RCA), Carry look –Ahead Adder, carry select adders, carry save adders and so on.

3.2 CMOS Adders Architectures: Ripple Carry Adder, Carry Look-Ahead Adder, Carry Select Adders, Carry Save Addres

Q14. Explain the basic operation of ripple carry adder and evaluate its performance.

Model Paper-1, Q7(a)

Ans:

Basic Theory and Operation: Figure (1) illustrates an n-bit Ripple Carry Adder, which is constructed using n single-bit full adders. A full adder is the fundamental building block of the ripple carry adder (RCA). Usually, the single-bit full adder used in RCA is a 10-T full adder standard cell. It is possible to extend the ripple carry adder endlessly to any number of bits by joining the carry-out of the preceding 1-bit full adder to the carry-in of the subsequent 1-bit full adder. The carry bit ripples across the chain of cascading full adders as shown in the diagram. It starts at a lower bit and moves to the next higher level full adder.

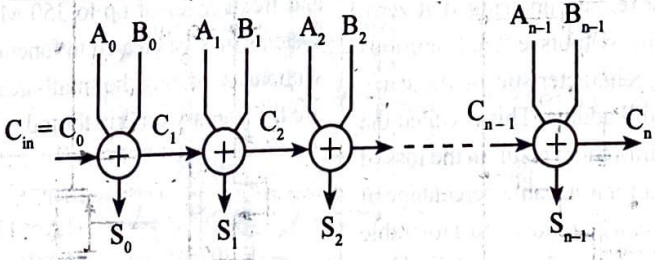


Figure (1): An N-Bit Ripple Carry Adder

Performance Evaluation: Among all of the adder designs, the RCA consumes the least amount of space and provides satisfactory performance for data that is entered randomly. But, its delay characteristics are directly proportional to the length of the carry propagating line. In other words, the delay of RCA increases with an increase in the carry propagating line. This makes it an almost poor choice for circuits that make use of nonrandom input operands. In RCA, a worst-case condition of addition occurs when the carry has to ripple all the way from the position of the LSB to that of the MSB. This is because all of the full adders in the RCA are connected to each other by the carry chain. The worst-case delay increases linearly with an increase in the number of bits processed by the operands as the length of the carry propagation line is proportional to the number of bits processed. Further, the space consumed by the RCA is also proportional to the number of bits processed. Due to this reason, in conditions where minimum amount of hardware is required and high speed performance is extremely important use of RCA for arithmetic operation will be counterproductive.

Q15. Explain the glitching problem in ripple carry adder.

Ans:

A glitch is a small spike that happens at the output of a digital circuit before the output reaches its steady state. Glitching power dissipation is the power consumed by the glitches.

The ripple carry adder (RCA) experiences difficulty with glitching in its system. To comprehend the issue, we consider a 4-bit ripple-carry adder that experiences glitching. Figure (1) illustrates the 4-bit ripple carry adder and Figure (2) depicts its static simulation.

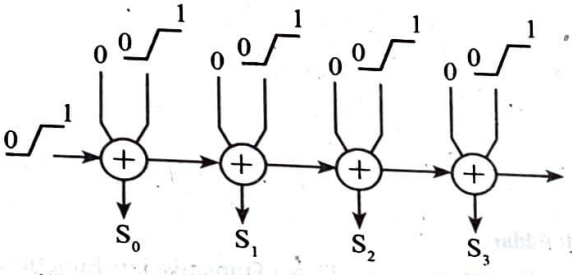


Figure (1): A 4-Bit Ripple Carry Adder Experiencing Glitching

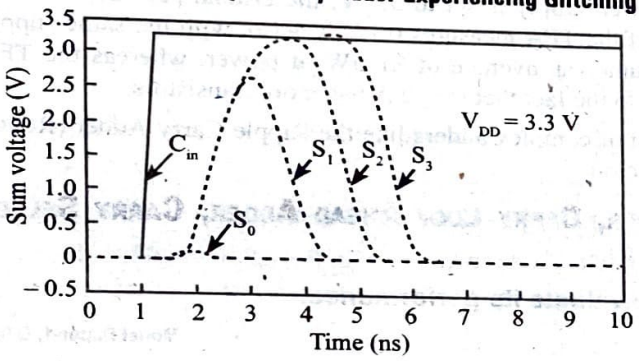


Figure (2): Static Simulation of 4-Bit Ripple Carry Adder

In this case, we simply assume that the inputs Ai are set to zero, while the input B_i and C_{in} both raise from 0 to 1. Under this conditions theoretically, the value of outputs S_{out} of the 4-bit ripple carry adder must remain unchanged at zero. But, as illustrated in figure (2), the outputs exhibit spurious (false) transitions due to the delay characteristic of the carry signal across its chain of cascaded full adders. This is called the glitching phenomenon. These dynamic shifts result in the loss of additional power, which can account for a sizeable percentage of the total power. Still, this drawback can be reduced to a tolerable level by carefully designing and optimizing the circuit.

Q16. With an example explain how carry propagation time can be minimized for ripple carry adder.

Ans: **Model Paper-3, Q6(b)**

The carry propagation time of ripple carry adder can be minimized by using various designs of improved full adder standard cell. The use of fastest full adder standard cell in the construction of ripple carry adder minimizes the carry propagation time. One illustration of such minimization is shown in the following example.

Consider two 32 bit ripple carry adder designs, one which is designed using the Transmission Gate Full Adder (TFA) model and the other is designed using the 10-transistor full adder (10-T FA) model. When the power supply voltage is 2.8 volts. The 32-bit ripple carry adder designed using the TFA model has a critical path delay of 7.2 ns. Whereas the 32-bit ripple carry adder designed using the 10-T FA model has a critical path delay of 4.1 ns. This means the 10-T FA model exhibits 44% more speed than the TFA model. The delay time as a function of the power supply voltage is illustrated in figure (1). In this the power supply ranges from 2.8 to 5V. It is very clear that the 10-T FA model performs better than the TFA model over the entire power supply range of 2.8 to 5V.

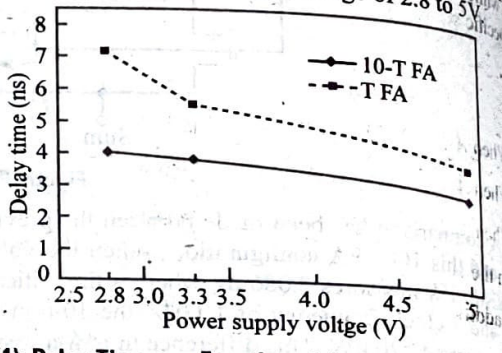


Figure (1): Delay Time as a Function of the Power Supply Voltage

From power consumption perspective, it is observed that the 10-T FA model dissipates 2.1 mW, whereas the TFA model dissipates 11 mW. This means the 10-T FA model consumes 81% less power than the TFA model. The simulation was run at a supply voltage of 2.8 volts and a clock frequency of 125 megahertz for both of the 32-bit ripple carry adders. Figure (2) illustrates the power consumption that occurs at various levels of supply voltage and clock frequency. It is clearly evident that the 10-T FA model consumes less power in comparison to the TFA model over the entire power supply range of 2.8 to 5V. The 10-T FA can function successfully at a supply voltage of 5 V and frequencies of up to 350 MHz. This indicates that massive systems may be created to function at extremely high frequencies without sacrificing the small-area and low-power qualities, which are the primary criteria for today's emerging technology.

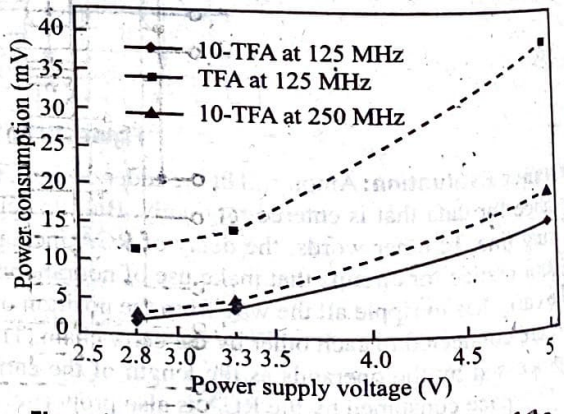


Figure (2): Power Consumption as a Function of the Power Supply Voltage

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Q17. Explain Basic theory and operation of carry look-ahead adder (CLA).

Ans:

In the case of a ripple carry adder, the size and delay associated with the logic circuit increase linearly with input operands. That is, as input increases, the size and delay of the adder increases. In order to overcome this drawback, the carry look-ahead adder has been developed. As the name indicates, in this circuit, the carry of each stage is produced in advance. That is, using look-ahead logic, carry is generated simultaneously (or in parallel) with the addition operation. As a consequence, it yields time propagation duration of $O\log(n)$ and a size requirement of $O(n\log n)$. So, the delay time of the CLA design displays a logarithmic dependence on the size of the adder. This enables the propagation delay of the carry signal to be reduced to the greatest extent possible.

In carry look-ahead adder, carry is expressed as function of the appropriate propagate and generate signals P_i and G_i, as well as the initial carry-in signal, C_{in}. The propagate signal P_i decides whether the carry-in to the ith block may propagate to the output, while the generating signal G_i decides if a carry-out would be set from within the block independently from the inputs. For a specific set of inputs A_i and B_i, the expressions for G_i and P_i are as follows:

$$G_i = A_j \bullet B_i \quad \text{... (1)}$$

$$P_i = A_i \oplus B_i \quad \text{... (2)}$$

When $A_i = B_i$ a carry gets generated

When $Ai = B_i = 1$, a carry of 1 is produced at the ith position, and

When $A_i = B_i = 0$, a carry of 0 gets generated in its place.

On the contrary, when $A_i \neq B_i$ carry propagation take place.

In addition to the P and G signals, the other Boolean variables of the Carry Look-Ahead adder are as follows:

$$S_i = P_i \oplus C_i \quad \text{... (3)}$$

$$C_{i+1} = G_i + P_i \bullet C_i \quad \text{... (4)}$$

Where "S_i" denotes the sum of the ith stage and "C_{i+1}" denotes the carry recurrence that occurs at ith stage.

Figure (1) illustrates the logic schematic of a 4-bit carry generator and figure (2) illustrates the block diagram of a 4-bit Carry Look-Ahead adder. From figure (1), it is evident that only two gate delays are required for carry generation. This enables the addition of two n-bit inputs much faster than the Ripple Carry Adder. On the other hand, to build this logic circuit a greater number of gates are required due to the fact that for high values of n, a massive amount of gates and a very large fan-in are essential.

For adders with lengths less than or equal to four, a Carry Look-Ahead adder is not effective that's because a four-stage Ripple Carry Adder needs less hardware and is virtually as fast. Also, the Carry Look-Ahead adder uses more space and power due to its huge number of logic gates.

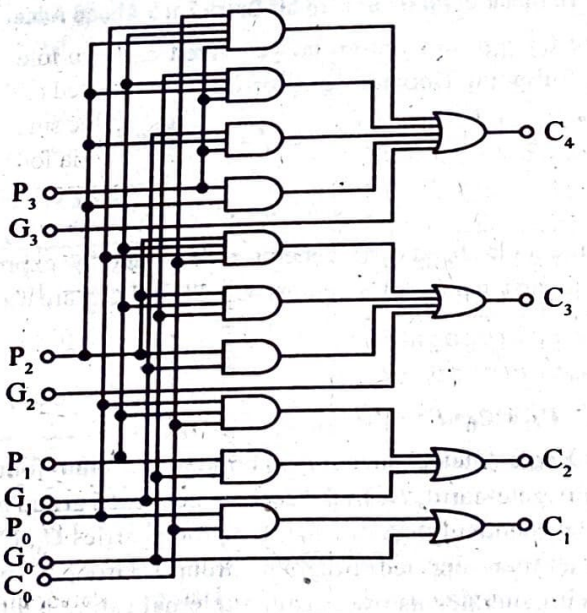


Figure (1): A 4-bit Carry Generator

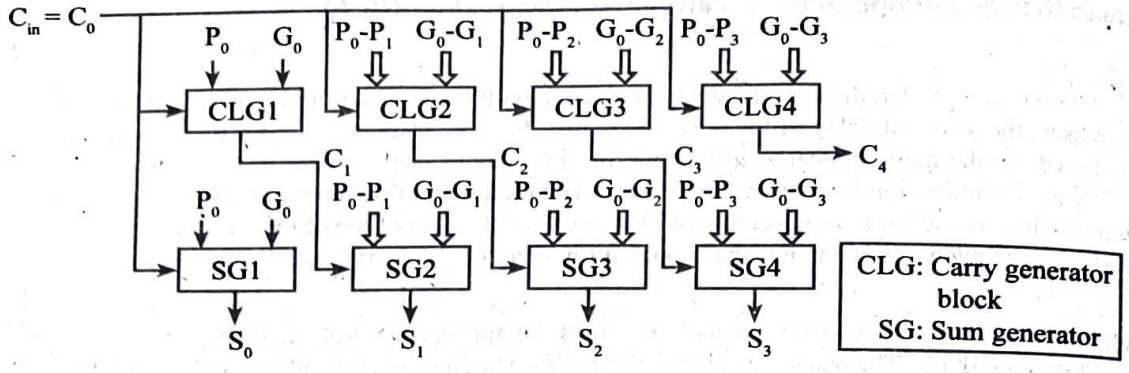


Figure (2): Block Diagram of a 4-bit Carry Look-Ahead Adder

Q18. Draw the block diagram of 16-bit Carry Look-Ahead adder and Explain.

Ans: Model Paper-3, Q7(b)

In the construction of the Carry Look-Ahead adder, for higher values of *n*-bit inputs, a massive number of gates and a very large fan-in are essential. Hence, to decrease the span of the carry look-ahead, the *n*-bit inputs are split into 4 equal groups, which are then cascaded in series by implementing the idea of Ripple Carry Adder. This technique undoubtedly minimizes the speed of carry propagation.

Figure (1) illustrates the block diagram of a 16-bit Carry Look-Ahead adder, which is split into four equal 4-bit groups. It also encompasses a combinatorial circuit, specifically the look-ahead carry generator.

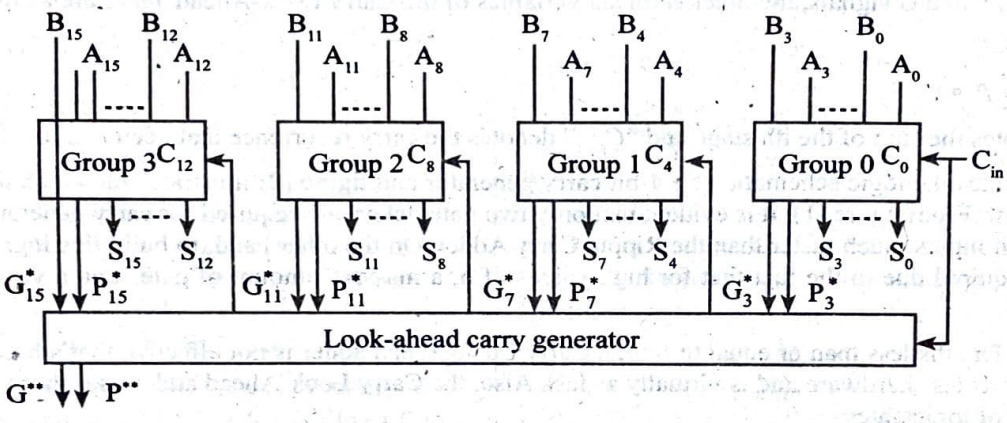


Figure (1): Block Diagram of a 16-bit Carry Look-Ahead Adder

In the block diagram the descriptor G_k^* signifies the group-generated carry, while P_k^* signifies the group-propagated carry. These carries are represented by the following Boolean equations:

$$G_k^* = G_k + G_{k-1} \cdot P_k + G_{k-2} \cdot P_{k-1} \cdot P_k + G_{k-3} \cdot P_{k-2} \cdot P_{k-1} \cdot P_k \quad \ldots (1)$$

$$P_k^* = P_{k-3} \cdot P_{k-2} \cdot P_{k-1} \cdot P_k \quad \ldots (2)$$

Where, $k = 3, 7, 11, 15$

These equations are executed using the look-ahead carry generator block, and the expressions of G_k^* and P_k^* are utilized to produce the group carry-in's. The look-ahead carry generator's outputs, C_4, C_8, and C_{12}, are used as inputs to the succeeding groups.

$$C_4 = G_3^* + C_0 \cdot P_3^* \quad \ldots (3)$$

$$C_8 = G_7^* + G_3^* \cdot P_7^* + C_0 \cdot P_3^* \cdot P_7^* \quad \ldots (4)$$

$$C_{12} = G_{11}^* + G_7^* \cdot P_{11}^* + G_3^* \cdot P_7^* \cdot P_{11}^* + C_0 \cdot P_3^* \cdot P_7^* \cdot P_{11}^* \quad \ldots (5)$$

The functioning of the 16-bit Carry Look-Ahead adder can be broken down into four distinct steps. To start, every group generates bit-generate-carry, G_i and bit-propagate-carry, P_i. In the second step, each group generates G_k^* and P_k^* (that is, group-generate-carry and group-propagate-carry, respectively). After that, the group carries C_4, C_8, and C_{12} are generated by the carry look-ahead generator. These group carries are then supplied straight to Group 1, Group 2, and Group 3 as illustrated in figure (1). In the last step, each of the four groups first generates its own distinct internal carry, and then proceeds to create the sum bits. Additionally, the section-carry generate (abbreviated as G^{**}) and section-carry propagate (abbreviated as P^{**}) are calculated.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Q19. Explain in brief about ELM adder.

Model Paper-1, Q6(b)

Ans:

An ELM adder is a variation of the basic Carry Look-Ahead adder algorithm. The algorithm of ELM addition encompasses a binary tree of basic processors that run in $O(\log n)$ time. The ELM algorithm was developed using the idea of carry propagate and generate. The ELM adder directly calculates the sum bits in parallel, which allows it to lower the number of inter-cell connections. This is the innovation of the ELM adder. The schematic diagram of an 8-bit ELM adder is shown in figure (1).

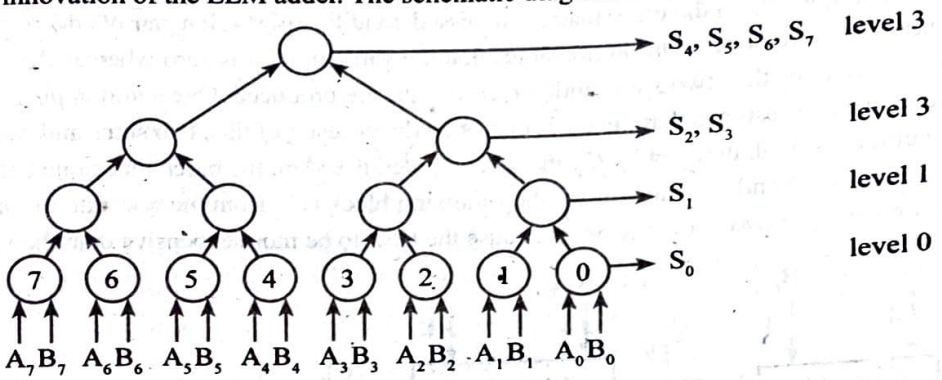


Figure (1): Block Diagram of a 8-bit ELM Adder

In the binary tree of the ELM algorithm, the leaves present at the lowest level will receive the binary inputs, A_i and B_i. At each level, the rightmost node calculates that level's output, which is denoted as S_i. Levels 0 and 1 both contribute one bit to the final sum, denoted by S_0 and S_1, respectively. At the same time, the rightmost nodes of the subsequent levels calculate 2^{k-1} bits; here k is the number of the level in the tree. The carry produce signals, carry propagate signals, and partial sums are computed by the tree's leaves. The tree's higher level nodes receive the partial sums generated from the adjacent lower level of the tree and also generate and propagate data necessary to process the partial sums at that specific level of the tree. Once the complete relevant data is available, new partial sums and the resulting carry generate and propagate signals get estimated and passed up to the subsequent higher level of the tree.

Q20. Compare and contrast the performance of CLA, ELM and RCA adders.

Ans:

Table (1) presents a concise understanding of the comparative aspects of the CLA, ELM and RCA adders. The static CMOS circuit design technology was used to simulate the 32-bit Carry Look-Ahead and ELM adders. From the simulation it is evident that each of these layouts needs more transistors and takes up more space than the Ripple Carry Adder. The study considers not just the transistors, it also considers the interconnections. Although the CLA adder has much more transistors than the ELM adder, it does have fewer interconnects and so consumes less space.

Adder type	Area ($\times 10^6 \lambda^2$)	Number of transistors	Delay (ns)	Average power dissipation per addition (mW)
CLA	2.27	2132	15	114.6
ELM	2.36	2078	10	104.1
RCA	0.80	1204	55	87.2

Table (1): Comparison of CLA, ELM and RCA adders

The ELM adder has the lowest worst-case delay, preceded by the CLA and RCA. In other words, in terms of worst-case delay the ELM adder out performs both the CLA and RCA. The RCA adder has the poorest delay performance. RCA undergoes the maximum delay due to the long carry chain. Meanwhile, the CLA consumes much more power as compared to its ELM and RCA counterparts.

Q21. Explain about carry select adder (CSL).

Ans:

The Ripple Carry Adder uses a small area and has a longer delay, whereas the Carry Look-Ahead adder uses a larger area and has a shorter delay. The Carry Select Adder (CSL) provides a considerable trade-off between both the RCA and CLA. Both the n-bit inputs, A_i and B_i that are used in the Carry Select Adder are split into k blocks of diverse sizes. Figure (1) illustrates a prototype of an 8-bit Carry Select adder that contains constant size blocks. The initial block with $C_{in} = C_o$ is realized by a 4-bit adder, which is usually the RCA. The second block, on the other hand, is assessed conditionally with a pair of adders, which permits two additions to be carried out in parallel. In this one addition presumes that the carry-in (C_{in}) is zero whereas the other addition presumes that the C_{in} is one. As a direct result of this, two sums and two carry-outs are produced. One addition presumes that the carry-in (C_{in}) is zero whereas the other addition presumes that the C_{in} is one. As a direct result of this, two sums and two carry-outs are produced. It is necessary to have a control signal, denoted by C_4, that is computed based on the block that came before it in order to choose the appropriate set of sum bits (S_{4-7}) and the carry-in for the following block (C_8) from the relevant multiplexers. The duplicate carry chain and the choose logic are the two components that cause the CSL to be more expensive than the RCA.

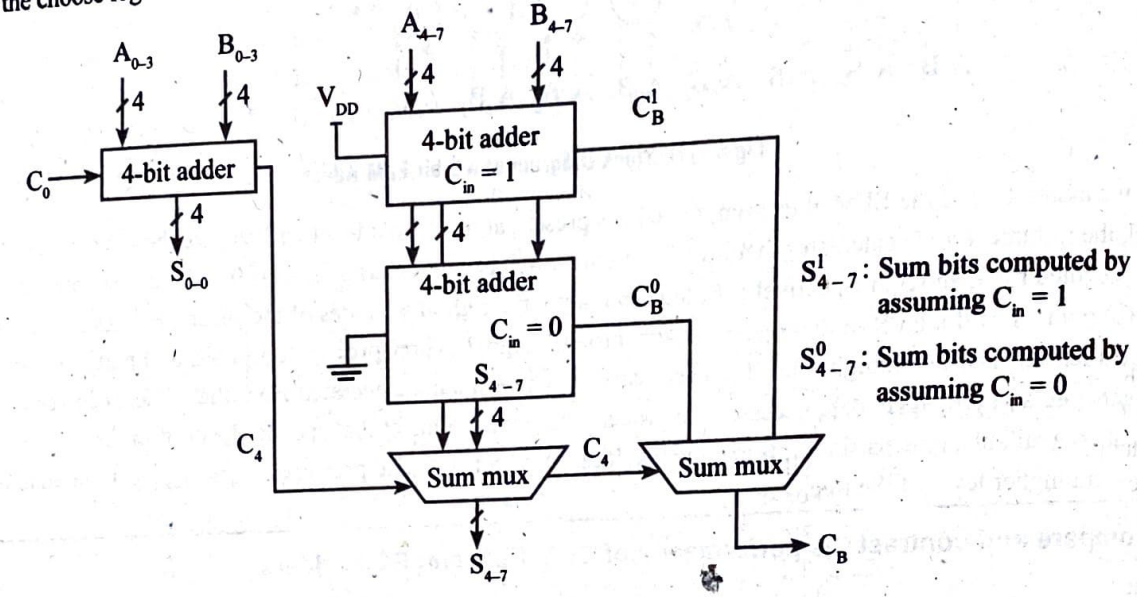


Figure (1): A Prototype of 8-bit Carry Select Adder

Performance

The RCA and CSL adders have their layouts designed for the following sizes: 8 bits, 16 bits, 32 bits, 64 bits, and 128 bits. Table below provides a summary of the comparisons that were made about the differences in area sizes and performance delays between the two kinds of adders. When compared to the RCA, it can be shown that the classical **Carry Select Adder** has a performance delay reduction of 59% on average; however this comes at the expense of using a much greater area.

Number of bits	Area, λ^2		%Change	Delay,ns		% Change
	RCA	CSL		RCA	CSL	
8	154624	290160	87.6	11.5	8.6	–26.0
16	382720	717889	87.5	25.0	11.5	–54.0
32	914400	1779904	94.6	52.5	21.5	–59.0
64	2439168	4667608	91.3	108.0	33.0	–69.4
128	7115072	13536432	90.2	226.0	54.0	–76.1
Average			90.2			–56.9

Table: CSL vs RCA Area and Delay Comparison

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Q22. Explain in brief about carry save adder (CSA). Model Paper-1, Q7(b)

Ans: Carry Save Adder (CSA) is in fact a Ripple Carry Adder (RCA) with its carries saved instead of propagated. In other words, Carry Save Adder is obtained by slight modification of the Ripple Carry Adder, that is, the carrier is saved within the full adder instead of propagating it to next full adder block. Figure (1) illustrates the block diagrams of both RCA and CSA for comparison. The RCA, CLA, and CSL adders are unable to perform concurrent addition of multiple inputs, which is a fundamental prerequisite for multiplication. But, the Carry Save Adder performs concurrent addition of multiple inputs.

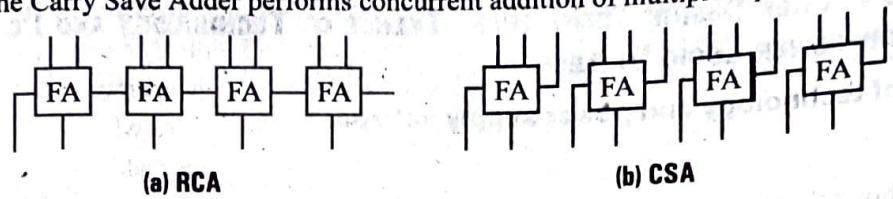


Figure (1): Block Diagram

A Carry Save Adder (CSA) tree is made using a CSA operator and one adder. The adder is located at the tree's root. Figure (2) illustrates the addition of three 1-bit binary values, *A*, *B* and *C*, with the CSA operator. The CSA operators are implemented in the addition process to convert an arbitrary number of operands into two adding operands, as a result of which the adder in the CSA tree evaluates the final sum. Because CSA saves the carry often it is called 3:2 counters.

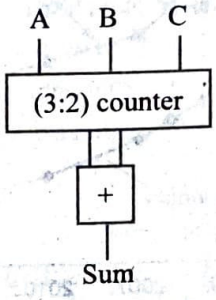


Figure (2): Addition Process with CSA Procedure

The cascading of CSA operators allows us to expand the 1-bit multi-operand addition to an *n*-bit multi-operand addition. An n-bit CSA is made up of '*n*' independent Full Adders working in parallel. Every Full Adder receives three ith bit inputs and produces two outputs. The two outputs are an ith bit partial sum (*S*) and an ith bit carry (*C*). If you want to add more than three operands, we may use a second or subsequent level of the CSA operators. They take the *S* and *C* from the preceding CSA operator level and add another input operand to create a new set of S and C values. There is no carry propagation in the levels with CSA operators. Only the final step propagates the carries. As a result, in contrast to the RCA's *n*-bit Full Adder propagation delay, the CSA has just one Full Adder propagation delay, which stays constant for any value of *n*.

Q23. Explain the addition of 4-bit binary numbers using carry save adder (CSA). Model Paper-2, Q7(b)

Ans: The addition of four 4-bit binary numbers *A*, *B*, *C*, and *D*, with an initial carry-in (C_0) using a Carry Save Adder (CSA) is illustrated in Figure (1). The top two rows of the FA or (3:2) counters make the 4-bit Carry Save Adder, whereas the third level forms the 4-bit Carry Propagating Adder (CPA). The CPA in this example is an RCA, but it may be replaced with a CLA or any other fast adder. Every CSA operator receives three bits that are of the equal relevance and then calculates the sum of the same relevance and the carry for the bit of a higher significance.

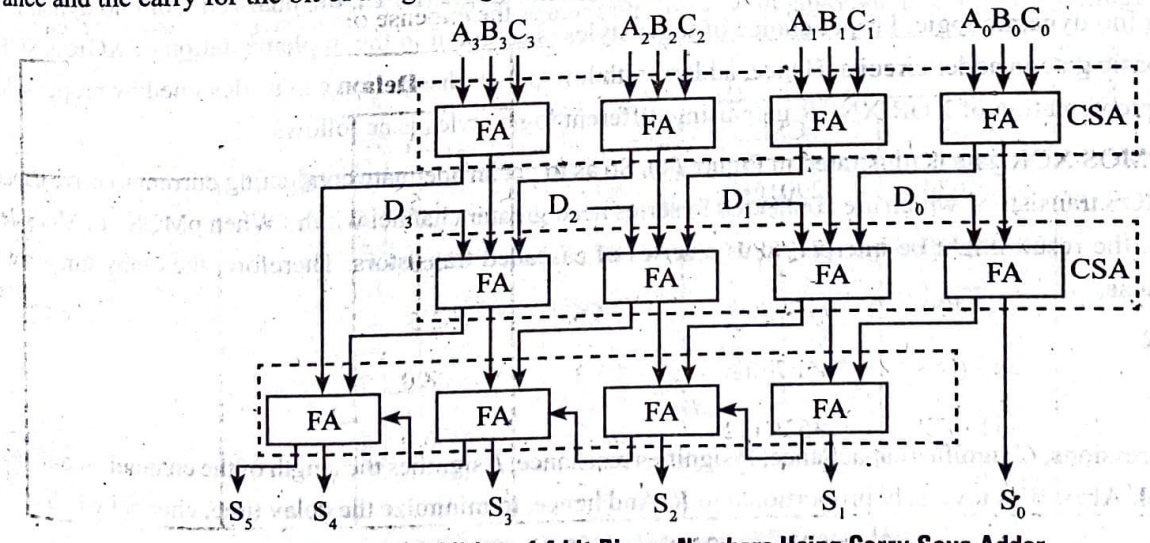


Figure (1): Addition of 4-bit Binary Numbers Using Carry Save Adder

The CSA's implementation may be extended to include k operands. To complete the addition operation $(k-2)$ CSA levels and one CPA are needed. The time it takes to get the summation is given as,

$$T = (k-2) \bullet T_{CSA} + T_{CPA} \quad \ldots (1)$$

Here, T_{CSA} and T_{CPA} are the executing times for both a CSA level and a CPA level, respectively.

3.3 Low-voltage Low-power Design Techniques –Trends of Technology and Power Supply Voltage, Low-voltage Low-power Logic Styles

Q24. Explain trends of technology and power supply voltage.

Ans:

The graph below shows the projected changes in printed gate length and power supply voltage for microprocessors (MPUs) after 2001 till 2016.

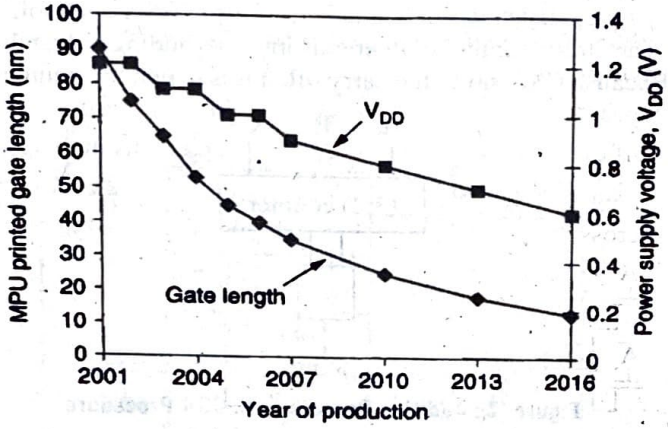


Figure: Trends of technology and power supply voltage

Scaled BiCMOS/CMOS technology is expected to continue to be the leading solution for low-voltage and low-power applications in the future. At 95 nm in 2001, the technology has lowered to 65 nm in 2003. It is possible that by 2016, the gate length will be reduced to 13 nm if manufacturing yield issues are resolved. In 2001, the power supply voltage was 1.2 V, and it is projected to fall to 0.9 V by 2007. In the long term, because of portability and reliability concerns, it is expected that the voltage will continue to fall to 0.6 V by 2016.

Q25. Explain the implementation of XOR/XNOR Gate using different logic styles.

Ans:

Model Paper-2, Q7(a)

The five different logic styles that are explicitly used for implementation of XOR/XNOR Gate are the full static CMOS, the Complementary Pass-transistor Logic (CPL), the Double Pass-transistor Logic (DPL), the dual-rail Domino dynamic logic, and the single-rail Domino dynamic logic. Proper choice of logic styles is essential in the implementation of XOR/XNOR gate because it is the most basic gate in adder circuits. Hence, adders with low-power dissipation can be designed by proper selection of logic styles. The implementation of XOR/XNOR gate using different logic styles is as follows.

The full static CMOS XOR gate is illustrated in Figure (1). So as to get an adequate conducting current to drive capacitive loads, the nMOS or pMOS transistors, which are connected in series need greater channel width. When pMOS or nMOS devices are connected in series, the result might be interpreted as a series of cascaded transistors. Therefore, the delay time that these devices impose is given as,

$$\tau = C.R \quad \ldots (1)$$

$$\frac{1}{R} \infty \frac{W}{L} \quad \ldots (2)$$

In the above expressions, C signifies capacitance, R signifies resistance, L signifies the length of the channel, and W signifies the width of the channel. Also, W is inversely proportional to R. And hence, to minimize the delay time, channel width (W) must be enhanced.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

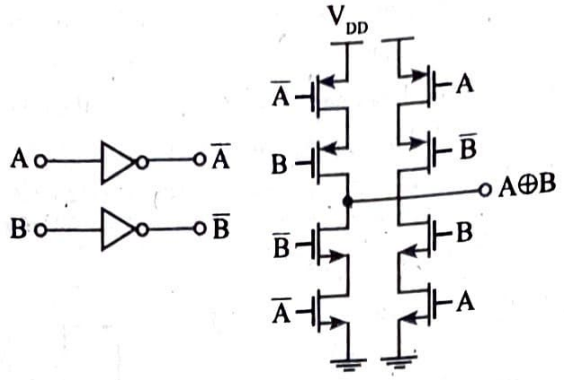


Figure (1): Full Static CMOS XOR Gate

Figure (2) is an example of an XOR/XNOR gate that makes use of the Complementary Pass-transistor Logic (CPL) in conjunction with a number of arrays of nMOS pass transistors. This logic style eradicates the issue of carefully sizing the series connected transistors, and hence this style needs only half the number of transistors as compared to the full static CMOS XOR gate. Whenever the output of the nMOS pass transistor network at node X is logically high, at $(V_{DD} - V_{th})$, where V_{th} is the threshold voltage, it induces an incomplete turn-off of the pMOS in the inverter, which results in a high short circuit current. This results in a major setback for the circuit, which causes a significant amount of loss. To limit the current and pull up the output node X to full V_{DD}, a pMOS device is connected across the inverter gate's output. This is depicted in figure (3).

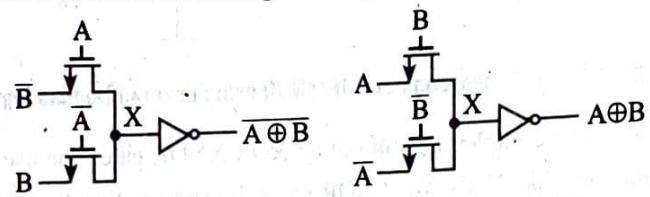


Figure (2): Implementation of XOR/XNOR using CPL

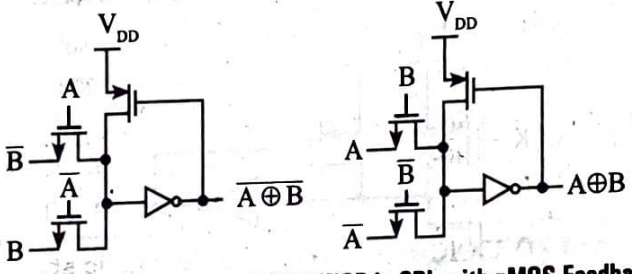


Figure (3): Implementation of XOR/XNOR in CPL with pMOS Feedback

The Double Pass-transistor Logic (DPL), which is a verification of the CPL, is yet another logic circuit that makes use of pass transistors. Figure (4) is an illustration of the XOR/XNOR gate that makes use of the DPL logic style. The issue of the nMOS threshold voltage drop, which occurs in the design of CPL logic, is avoided by the DPL since it makes use of both pMOS and nMOS devices.

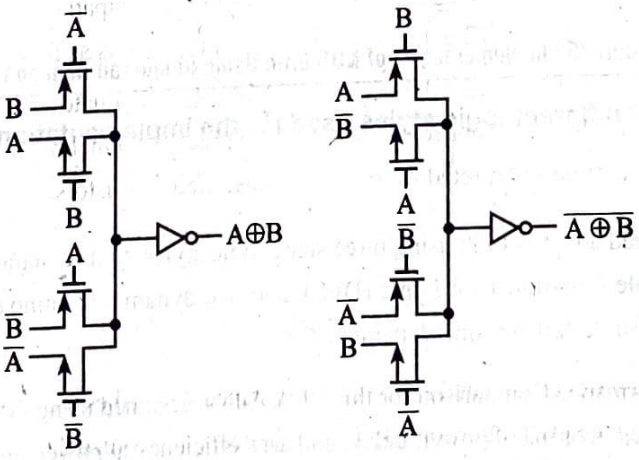


Figure (4): Implementation of XOR/XNOR using DPL

The implementation of XOR/XNOR gate using dual-rail Domino logic is seen in figure (5). In contrast to static system dynamic architecture begins with a precharge phase and end with an evaluation phase. When the CLK signal has a low value, precharge phase takes place, and when the CLK signal has a high value, the evaluation phase takes place. The dynamic design eliminates any spurious transitions and the related power consumption, which is inherently prevalent in any static logic design. This is made possible by the precharge and evaluation phases of the dynamic design. However, other components, like the clock driver, consume additional power.

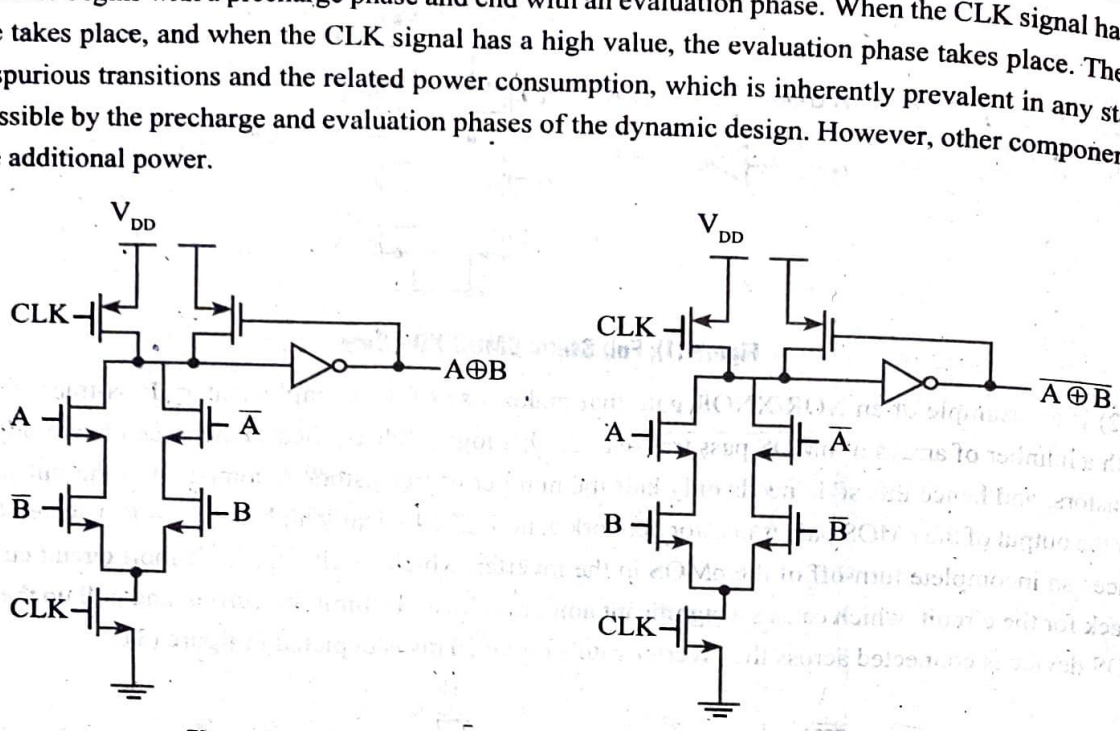


Figure (5): Implementation of XOR/XNOR using dual-rail Domino logic

The dual-rail is required only for the implementation of an XOR/XNOR gate. The use of single-rail logic is suitable for a variety of different logic gate functions, like NAND and NOR gates. Because of this, there is a reduction in the loss of power. Figure (6) illustrates implementation of the NOR gate using the single-rail Domino logic.

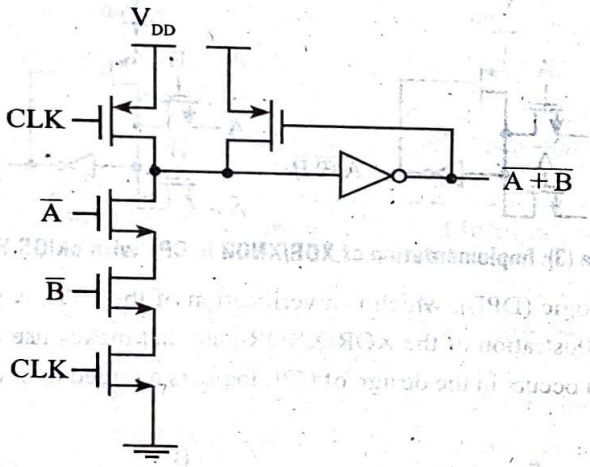


Figure (6): Implementation of NOR Gate Using Single-rail Domino Logic

Q26. Compare and contrast different logic styles used for the implementation of XOR/XNOR Gate.

Ans:

Model Paper-3, Q7(a)

A 32-bit Carry Look- Ahead adder is built using three static logic styles, that is, static CMOS, the Complementary Pass-transistor Logic (CPL), the Double Pass-transistor Logic (DPL), and two dynamic Domino circuit families, which are dual-rail Domino dynamic logic, and the single-rail Domino dynamic logic.

Table (1) provides a Performance Comparison for this CLA Adder designed using different logic styles. The comparison and evaluation of these logic styles in terms of power, delay, and area efficiency at power supply voltages ranging from 1.5 V to 3.5 V and an input clock frequency of 100 MHz is as follows.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Parameters	Full static CMOS	CPL	DPL	Dual-rail domino	Single-rail domion
Power (mW)	34.3	34.5	27.5	82.5	60.2
Worst case delay (ns)	2.33	2.24	1.98	1.78	1.64
Energy (pj)	79.9	77.3	54.5	146.9	98.7
Area total transistor width (μm)	27355	17437	19117	32444	26926

Table (1): Performance Comparison of CLA Adder Designed Using Different Logic Styles

The CLA adder designed using dual-rail Domino logic, consumes three times more power compared to the DPL adder, but it is still only 12 percent faster. When the CLA adder is designed using single-rail Domino logic, it consumes two times more power than the DPL adder and exhibits 21 percent faster speed. The energy efficiency of the static styles is much higher than that of the dynamic styles. To be more specific, the DPL adder is the most space- and resource-efficient of all the adders that have been realized using static circuit designs. It occupies 47 percent less area than the complete static CMOS adder and just 10 percent more area than the CPL adder.

IMPORTANT QUESTIONS

Q1. What is standard adder cell? Explain about standard half adder cell.

Ans: Refer Q11.

Important Questi

Q2. Explain various schematic configurations of Full adders their advantages and disadvantages.

Ans: Refer Q12.

Important Questi

Q3. Explain the basic operation of ripple carry adder and evaluate its performance.

Ans: Refer Q14.

Important Questi

Q4. With an example explain how carry propagation time can be minimized for ripple carry adder.

Ans: Refer Q16.

Important Questi

Q5. Explain Basic theory and operation of carry look-ahead adder (CLA).

Ans: Refer Q17.

Important Questi

Q6. Explain in brief about ELM adder.

Ans: Refer Q19.

Important Questi

Q7. Explain about carry select adder (CSL).

Ans: Refer Q21.

Important Questi

Q8. Explain in brief about carry save adder (CSA).

Ans: Refer Q22.

Important Questi

Q9. Explain the addition of 4-bit binary numbers using carry save adder (CSA).

Ans: Refer Q23.

Important Questi

Q10. Explain the implementation of XOR/XNOR Gate using different logic styles.

Ans: Refer Q25.

Important Questi

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

UNIT 4

LOW VOLTAGE LOWER POWER MULTIPLIERS

Look for the Logo on the Title Cover Before You Buy
SIA GROUP
SIA GROUP

Syllabus

LOW-VOLTAGE LOW-POWER MULTIPLIERS

Introduction, Overview of Multiplication, Types of Multiplier Architectures, Braun Multiplier, Baugh-Wooley Multiplier, Booth Multiplier, Introduction to Wallace Tree Multiplier.

LEARNING OBJECTIVES

- Concept of multiplication
- Different types of multiplier architectures
- Operation of Braun multiplier and its modified version
- Operation of Baugh-Wooley multiplier and its modified version
- Concept of Booth's algorithm
- Operation of Booth multipliers and its modified version
- Function of Booth encoder
- Working of Wallace multiplier
- Introduction to wallace tree multiplier
- Wallace three construction

INTRODUCTION

In arithmetic operations, multiplication is a fundamental function. Multiplication-based operations like Multiply and Accumulate (MAC) and inner product are among the most commonly utilized computation-intensive arithmetic functions in many Digital Signal Processing (DSP) applications like convolution, Fast Fourier Transform (FFT), filtering, and others. In this unit, multiplication operation and different types of multiplies are discussed.

PART-A SHORT QUESTIONS WITH SOLUTIONS

Q1. What are the variables to be considered while designing a multiplier.

Ans:

The three significant variables to considered while designing multipliers are,

1. Chip area.
2. Calculation speed.
3. Per dissipation.

Q2. What are the different types of multiplier architectures. Model Paper-I

Ans:

The different types of multiplier architectures are,

1. Serial multiplier
2. Parallel multiplier
3. Serial-parallel multiplier.

Q3. Mention the various multiplier architectures followed for VLSI design.

Ans:

The different types of multiplier architecture followed in VLSI design are as follows,

1. Serial parallel multiplier
2. Braun multiplier
3. Baugh-Wooley multiplier
4. Array multiplier
5. Wallace tree multiplier.

Q4. What is serial parallel multiplier?

Ans:

Serial-parallel Multiplier is one of the simplest available multiplier. The operation of this multiplier is explained bel

Let, $A = (a_n, a_{n-1}\ldots\ldots, a_0)$

$B = (b_n, b_{n-1}\ldots\ldots, b_0)$

Then the multiplication of A and B is given by,

$$A.B = (A.2^n.b_n + A.2^{n-1}\, b_{n-1} + \ldots\ldots + A.2^0.b_0)$$

Q5. What is Braun multiplier?

Ans:

Braun Multiplier this type of multiplier, all the partial products are computed in parallel and collected using a cas array of carry save adders.

Q6. What is Baugh-Wooley multiplier? Model Paper-I

Ans:

Baugh-Wooley Multiplier is used for two's complement multiplication. It is explained below.

$$\text{Let, } A = (a_{n-1}\ldots\ldots a_0) = -a_{n-1}\, 2^{n-1} + \sum_{0}^{n-2} a_i.2^i$$

$$B = (b_{n-1}\ldots\ldots b_0) = -b_{n-1}\, 2^{n-1} + \sum_{0}^{n-2} b_i.2^i$$

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Then, the multiplication of A and B is given by,

$$A.B = a_{n-1}\, b_{n-1}\, 2^{n-2} + \sum_{0}^{n-2}\sum_{0}^{n-2} a_i b_i 2_i - a_{n-1}\sum_{0}^{n-2} b_i 2^{n+i-1} - b_{n-1}\sum_{0}^{n-2} a_i 2^{n+i-1}$$

In order to avoid subtraction, the negative terms are interchanged by using appropriate simplification.

Q7. What is Pipelined multiplier and wallace Tree Multiplier? **Model Paper-2, Q1(i)**

Ans:

Pipelined Multiplier

- Pipelined Multiplier is the multiplier using latches, we can obtain a pipelined multiplier. These type of multipliers are used in case where no-storage elements are present.
- Let the multiplier be Y, multiplicand be X and the product be P.
- Then, the k^{th} bit of each partial product $(X_{k-i} \times Y_i)$ is formed as,

$$P_k = \sum_{i=0}^{k} X_{k-i} \times Y_i$$

Wallace Tree Multiplier

- These multipliers are named after their inventor. These multipliers can be constructed from adder cells.
- The main concept of this multipliers revolves around designing multipliers, whose completion time grows as the logarithm of the number of bits that are to be multiplied.

Q8. Explain the significance of Booth's algorithm. **Model Paper-3, Q1(i)**

Ans:

Booth's algorithm is the strongest and effective algorithm for solving signed 2's complement multiplication. It produces 2 n-bit product and gives equal priority to both positive and negative numbers. This algorithm adopts a technique in which multiplication is represented as difference between numbers, thereby reducing the number of operations required for multiplication. In this algorithm, 0's in the multiplier require just shifting and 1's in the multiplier from bit position n to bit position m can be treated as $2n + 1 - 2m$.

Q9. State radix-2 booth encoding table. **Model Paper-2, Q1(d)**

Ans:

Radix-2 booth encoding table is shown below.

Y_i	V_{i-1}	Partial Product
0	0	0 × Multiplicand
0	1	1 × Multiplicand
1	0	–1 × Multiplicand
1	1	0 × Multiplicand

Q10. What are the steps involved in a Wallace tree?

Ans:

The Wallace tree has the following three steps.

1. Partial Product Generation Stage
2. Partial Product Reduction Stage
3. Partial Product Addition Stage

Q11. What is a 4:2 compressor? **Model Paper-3, Q1(d)**

Ans:

A 4:2 compressor is a combination of two full adders. It has four input bits In_1, In_2, In_3 and In_4 and two output bits sum and carry (S and C) along with an input carry bit , C_{in} and output carry bit, C_{out}.

PART-B ESSAY QUESTIONS WITH SOLUTIONS

4.1 Introduction, Overview of Multiplication

Q12. Give an introduction to multiplication in brief.

Ans:

In arithmetic operations, multiplication is a fundamental function. Multiplication-based operations like Multiply and Accumulate (MAC) and inner product are among the most commonly utilized computation-intensive arithmetic functions in many Digital Signal Processing (DSP) applications like convolution, Fast Fourier Transform (FFT), filtering, and others.

They frequently generate a large amount of time delay and consume a significant amount of silicon area in the DSP system. Because multiplication takes up the majority of the execution time in most DSP algorithms, a high-speed multiplier is essential. Multiplication time is still the most important component in determining a DSP chip's instruction cycle time.

With the growing demand for more computational power on battery-powered mobile devices, the focus has changed from optimizing traditional delay times and area sizes to minimizing power waste while maintaining good performance. A low-power design, in comparison to prior designs, allows portable devices to work for longer periods of time with the same amount of battery charge.

Q13. Discuss the concept of multiplication with an example.

Ans: Model Paper-1, Q8(a)

Multiplier is an arithmetic operator which performs multiplication between two binary numbers and generates the result. For $M \times N$ bit multiplication, N partial products are formed for each M bit and then the shifted partial products are added to generate $M \times N$ bit result. The operation of binary multiplication is similar to logical AND operation, So, each partial product includes, ANDing of multiplicand and multiplier bits.

Consider $B = (b_{M-1}, b_{M-2}, b_{M-3}, \ldots b_1, b_0)$ as multiplicand and $A = (a_{N-1}, a_{N-2}, a_{N-3}, \ldots a_1, a_0)$ as multiplier. Then, the product of multiplicand and multiplier bits is given,

$$A = \left[\sum_{j=0}^{M-1} b_j 2^j\right]\left[\sum_{i=0}^{N-1} a_j 2^i\right]$$

$$B = \sum_{i=0}^{N-1}\sum_{j=0}^{M-1} a_i b_j 2^{i+j}$$

Binary multiplication is illustrated in the following example.

Multiplicand, $13_{10} = 1\ 1\ 0\ 1$

Multiplier, $6_{10} = 0\ 1\ 1\ 0$

$$\begin{array}{rcccccccc} & & & & 1 & 1 & 0 & 1 & \\ & & & \times & 0 & 1 & 1 & 0 & \\ \hline & & & & 0 & 0 & 0 & 0 & \\ & & & 1 & 1 & 0 & 1 & & \\ & & \textcircled{1} & & & & & & \\ & & 1 & 1 & 0 & 1 & & & \\ & \textcircled{1} & & & & & & & \\ 0 & 0 & 0 & 0 & & & & & \\ \hline (78)_{10} = & 1 & 0 & 0 & 1 & 1 & 1 & 0 & \end{array}$$

Partial products

Products

$\therefore \quad 13_{10} \times 6_{10} = 78_{10}$

Q14. Give an overview of multiplication.

Ans: Model Paper-2, Q8

Multiplication can be viewed as a series of repeated additions that are repeated. The multiplicand is the number to be multiplied, the multiplier is the number of times it is added and the result is the product.

Multiplication involves the basic actions of producing and accumulating or adding partial products. As a result, the two primary phases must be streamlined to speed up the entire multiplication process.

1. Unsigned multiplication
2. Signed multiplication

1. Unsigned Multiplication

Unsigned multiplication can be performed on a processor using AND gates and full adders.

This multiplication involves the following two steps

- Each digit of multiplier is multiplied by the multiplicand for partial result
- The partial results are added to obtain final result

For instance, consider two n-bit unsigned numbers A and B defined by,

$$A = \sum_{i=0}^{n-1} A_i 2^i$$

$$B = \sum_{j=0}^{n-1} B_i 2^j$$

Where,

A - Multiplicand

B - Multiplier

Then, the product of A and B is given by,

$$P = \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} A_i B_j 2^{(i+j)}$$

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

They can be expanded as,

A Multiplicand $A_{n-1}A_{n-2}A_{n-3}\ldots\ldots A_1 A_0$

B Multiplier $B_{n-1}B_{n-2}B_{n-3}\ldots\ldots B_1 B_0$

P Product $(A*B)\ P_{2n-1}P_{2n-2}P_{2n-3}\ldots\ldots P_1 P_0$

Each of the partial product terms $P_n = A_i A_j$ is called the Summand.

Each partial product is then stored in the Arithmetic Logic Unit register, where it takes up memory until the final partial product is calculated. The final result is obtained by adding all of the partial products.

Signed Multiplication

Multiplication of signed numbers is performed in 2's complements form. Here, the negative number is initially converted to its 2's complements form to ensure that all partial products are positive.

One positive and one negative numbers

Consider two n-bit numbers A and $-B$. The negative number B is converted into represented by $(2^n - B)$.

The product on direct multiplication is then given by,

$$P' = A(-B) = A(2^n - B) = 2^n A - AB$$

This product differs from the expected result,

$$P = -AB = 2^{2n} - AB$$

where, $2^n(2^n - A)$ is the correction factor to be added. This is done by shifting the 2's complement of A to the left by n places. The pre-calculated product of P' is then added to this correction factor.

Two negative numbers

If both the multiplicand and multiplier are negative, their 2's complements will be multiplied.

Consider two n-bit negative numbers $-A$ and $-B$ in which $-A$ is to be multiplied with $-B$.

The product P' is given by,

$$P' = (2^n - A)(2^n - B) = 2^{2n} - 2^{2n}A - 2^n B + AB$$

It differs from the expect result,

$$P = AB$$

The difference

$$P - P' = -2^{2n} + 2^n A + 2^n B$$

The term 2^{2n} denotes a carry-out bit from the Most Significant Bit (MSB) that can be ignored. Correction factors for both multiplier and multiplicand must be included to get the correct result.

4.2 Types of Multiplier Architectures

Q15. What are the different types of multiplier architectures? Explain them.

Ans:

The different types of multiplier architectures are,

1. Serial multiplier
2. Parallel multiplier
3. Serial-parallel multiplier.

1. Serial Multiplier

- A successive addition algorithm is used by the serial multiplier.
- They have a simple structure because both operands are entered sequentially. As a result, the physical circuit uses less hardware and takes up less chip area.
- The serial multiplier, on the other hand, has a slow speed due to the operands being entered sequentially.

2. Parallel Multiplier

- To perform high-speed mathematical calculations, most advanced digital computers include a parallel multiplier. In its arithmetic logic unit, a microprocessor needs multipliers, and a digital signal processing system needs multipliers to implement techniques like convolution and filtering.
- In Reduced Instruction Set Computers (RISC), Digital Signal Processing (DSP), and graphics accelerators, high-speed parallel multipliers with substantially bigger areas and more complexity are widely used.

The examples of parallel multipliers are,

- Array multipliers such as the Braun and Baugh-Wooley multipliers
- Tree multipliers like the Wallace multiplier

Disadvantages

- The main disadvantage of these multipliers is that they need a lot of chip space.
- It has a fast processing speed, but it is costly in terms of silicon area and power consumption. This is because both operands are sent into the multiplier in parallel for parallel multipliers. As a result, compared to serial multipliers, the hardware takes up a lot more area and is more complicated.

3. Serial-Parallel Multiplier

A serial-parallel multiplier is one of the easiest and minutest methods to carry out the fundamental multiplication algorithm. It is a better trade-off between the time-consuming serial multiplier and the area-consuming parallel multipliers, which take up a lot of time and space. This type of multiplier is used when there is a need for both speed and a small area.

One operand is entered sequentially while the other operand is stored them in parallel with a fixed number of bits. A combinational full adder and a register are incorporated by the sum units in the multiplier that holds the carry. The shift and add operations are performed by the chain of summation units and registers. The partial product which is obtained from the above operation is held in the shift register chain, while the multiplicand is successively added into the partial product.

4.3 Braun Multiplier

Q16. What is Braun multiplier? Explain its operation.

Ans:

Braun multiplier

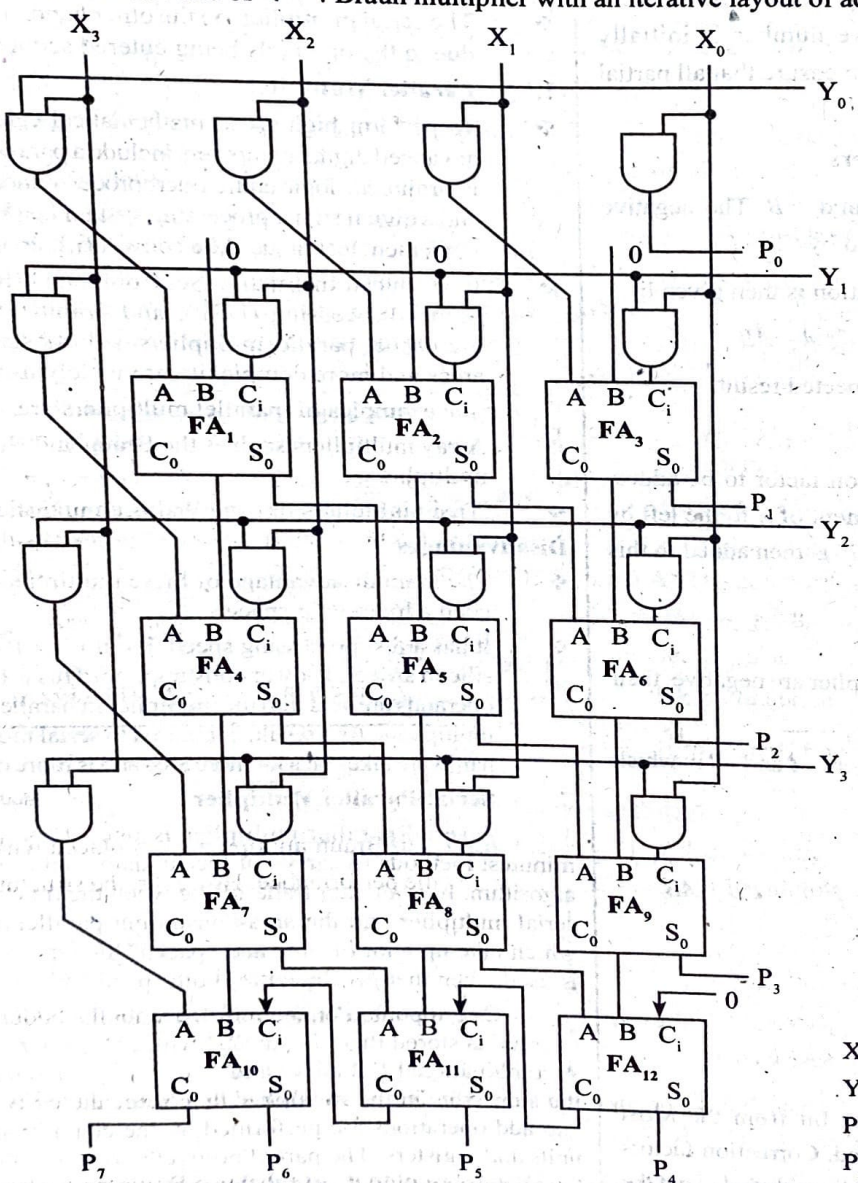
Model Paper

It is a basic parallel multiplier. It is also known as the Carry-Save Array Multiplier. This multiplier can only pe multiplication between two unsigned values. It is made up of an array of AND gates and adders that are placed in an ite structure without the need of logic registers. Because it does not add an additional operand to the multiplication result, it known as the non-additive multiplier.

Architecture

In Braun multiplier, all the partial products are computed in parallel and collected using a cascaded array of carr adders. An $n \times n$-bit Braun multiplier requires $n(n-1)$ adders and n^2 AND gates.

The efficient architecture of 4×4 Braun multiplier with an iterative layout of adders is as shown in figure (1).



X: 4-bit multiplicand
Y: 4-bit multiplier
P: 8-bit product of X and Y
P_n: X_iY_j is a product bit

Figure (1): Schematic Diagram of a 4 × 4-Bit Braun Multiplier

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Operation

- The AND gates generate each of the $X_i Y_j$ product bits in parallel.
- Using a row of adders, each partial product can be added to the previous sum of partial products.
- The carry-out signals are shifted one bit to the left before being added to the first adder's and the new partial product's sums.
- A Carry Save Adder is responsible for relocating the carry-out parts to the left (CSA).
- There is no horizontal carry propagation for the first four rows because the carry bits are transferred diagonally downward to the next adder stage. Instead, the corresponding carry bit is saved for the adder stage that follows.
- To output the final result, ripple carry adders (RCA) are utilized at the end stage of the array.

Speed Consideration

The Braun multiplier's delay is determined by the entire adder cell's delay as well as the final adder in the last row. Because the sum and carry signals are both in the critical path, a complete adder with balanced carry and sum delays is ideal in the multiplier array. For large arrays, the complete adder's speed and power are critical.

The worst-case multiplication time of a Braun multiplier can be expressed as,

$$t_{Braun} = (n-1)\, t_{carry\text{-}save} + t_{AND} + (n-1)\, t_{Ripple\text{-}Carry}$$

Where,

$t_{carry\text{-}save}$ - Time required to generate Carry-out (C_{out}) or Sum (S_{out}) at the output after the inputs are supplied to a CSA.

$t_{Ripple\text{-}carry}$ - Time taken for the Carry-out (G_{out}) or Sum (S_{out}) to be generated at the output after the inputs are supplied to a RCA

t_{AND} - Delay of an AND gate.

Q17. Give the schematic for a 4×4 carry-save multiplier and explain its operation.

Ans:

Figure shows a 4 × 4 carry-save multiplier.

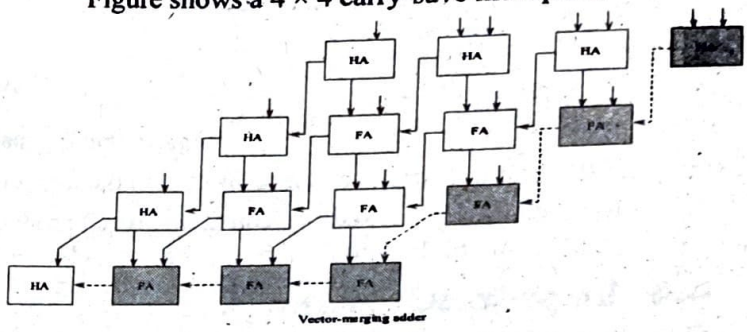


Figure: A 4×4 Carry-Save Multiplier (The Critical Path is Highlighted in Gray)

In a 4 × 4 carry-save multiplier, the output carry bits move downwards diagonally. This critical path can be observed as highlighted in figure. The final result is generated by an additional adder introduced known as vector-merging-adder. The carry-bits in this multiplier are not added instantly but are "saved" for the next adder stage and hence it is termed as the "Carry Save Multiplier".

The final stage of the multiplier combines both the carries and sums in a fast carry-propagate adder stage.

As the carry-save multiplier requires an additional adder, its area cost is increased a little. Because of this, the worst - case critical path is made short and is represented using the following expression.

$$t_{mult} = t_{add} + (N-1)\, t_{carry} + t_{merge}$$

Where,

t_{add} is considered to be equal to t_{carry} i.e., $t_{add} = t_{carry}$

Q18. Explain modified or enhanced Braun multiplier.

Ans:

Model Paper-1, Q8(b)

In enhanced Braun multiplier, the full adders FA_1, FA_2, FA_3, and FA_{12} of Braun multiplier are replaced with half-adders. Each replacement will result in a savings of three logic gates thus enhancing the performance. However, the structure's regularity is deformed.

Optimization Measures

- The Braun multiplier's performance can be improved by optimizing the interconnections between the adders, so that the delay across each adder's path is nearly equal.
- The connections are made on the basis that the previous adder's long delay path should be connected to the next adder's short delay path.
- Using two optimally interconnecting full adders with fast input and fast output, an enhanced Braun multiplier is shown in figure.

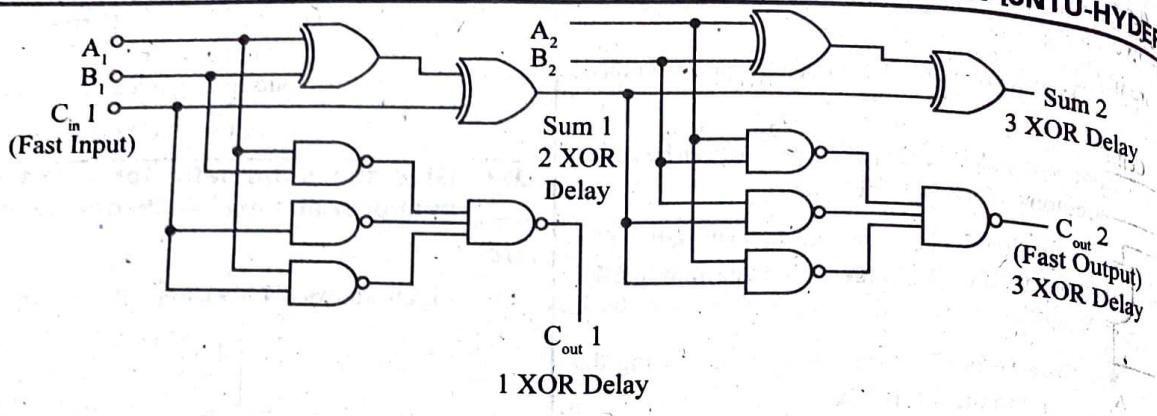


Figure: Two Optimally Interconnected Full Adders

- Either input A_2 or B_2 can be connected to the long delay path from the previous adder.
- Full adders can be used to connect the partial product array. Due to their logical similarity, any and all partial produ in the same bit location can be used alternately.

4.4 Baugh-Wooley Multiplier

Q19. What is Baugh-Wooley multiplier? Draw its structure.

Ans:

This multiplier is used for both unsigned and signed numbers with two's complement multiplication to ensure po results are positive.

The architecture of the Baugh-Wooley multiplier is based on the carry-save algorithm. It has a regular and iterative struc of the array multiplier. Figure (1) shows the structure of a 4 × 4-bit 2's complement multiplier, with the cell number indic the type of basic cell.

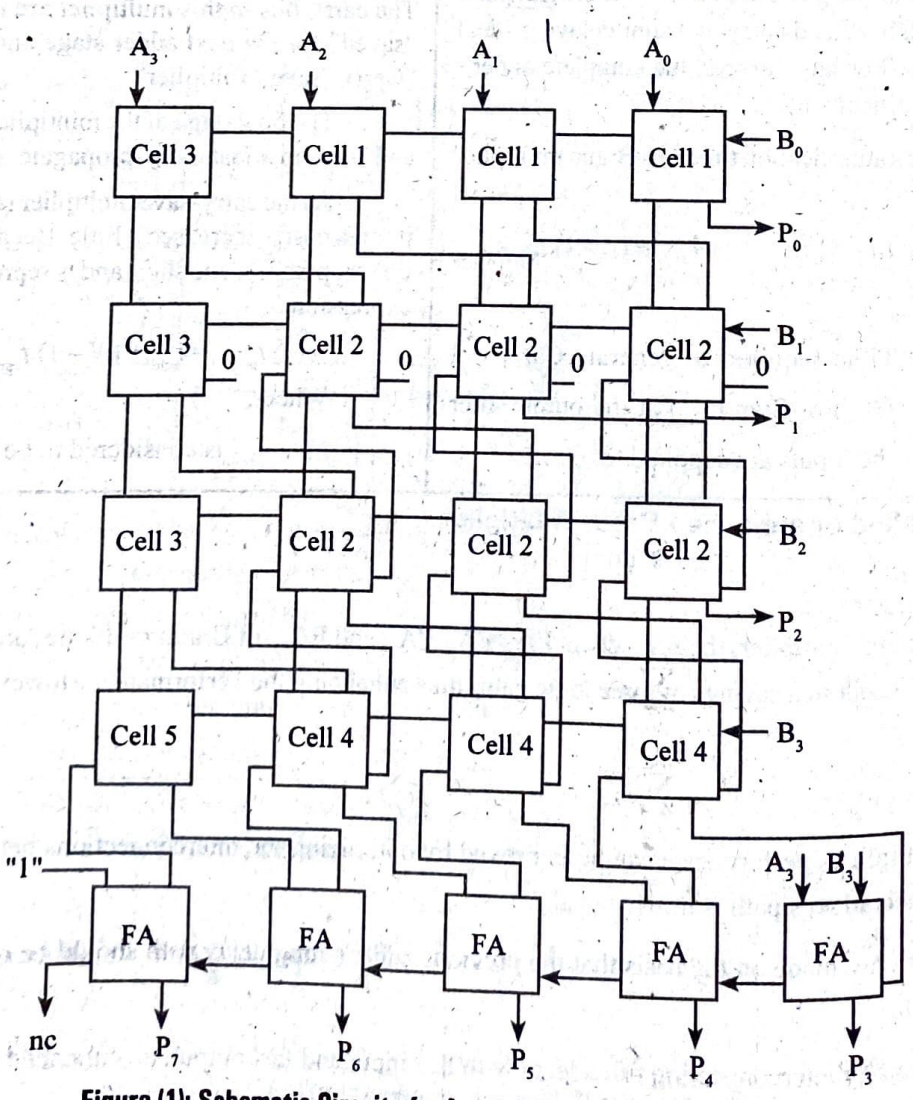


Figure (1): Schematic Circuit of a 4 × 4-Bit Baugh-Wooley Multiplier

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

The five basic building blocks of the Baugh-Wooley ultiplier are illustrated in figure (2).

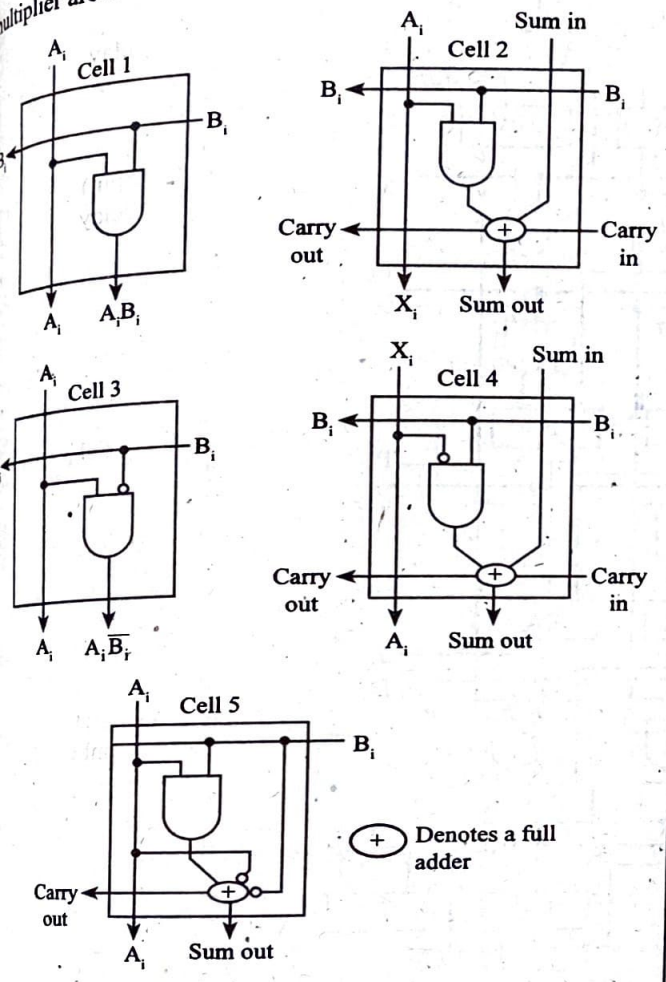


Figure (2): Basic Building blocks of Baugh-Wooley Multiplier

20. Explain Baugh-Wooley algorithm.

ns: **Model Paper-2, Q8(b)**

ugh-Wooley Multiplier

This multiplier is used for two's complement ultiplication. It is explained below.

Let, $A = (a_{n-1} \ldots\ldots a_0) = -a_{n-1}\, 2^{n-1} + \sum_{0}^{n-2} a_i 2^i$

$B = (b_{n-1} \ldots\ldots b_0) = -b_{n-1}\, 2^{n-1} + \sum_{0}^{n-2} b_i 2^i$

Then, the multiplication of A and B is given by,

$$A.B = a_{n-1}\, b_{n-1}\, 2^{n-2} + \sum_{0}^{n-2}\sum_{0}^{n-2} a_i b_i 2_i -$$

$$a_{n-1}\sum_{0}^{n-2} b_i 2^{n+i-1} - b_{n-1}\sum_{0}^{n-2} a_i 2^{n+i-1}$$

In order to avoid subtraction, the negative terms are interchanged by using appropriate simplification.

Therefore,

$$-A_{n-1}\sum_{j=0}^{n-2} B_j 2^{n+j-1} = A_{n-1}\left(-2^{2n-2} + 2^{n-1} + \sum_{j=0}^{n-2} \overline{B}_j 2^{n+j+1}\right)$$

The product is then given by,

$$P = AB$$

$$= -2^{2n-1} + (\overline{A_{n-1}} + \overline{B_{n-1}} + A_{n-1}B_{n-1})2^{2n-2} +$$

$$\sum_{i=0}^{n-2}\sum_{j=0}^{n-2} A_i.B_j 2^{i+j} + (A_{n-1} + B_{n-1})2^{n-1} +$$

$$A_{n-1}\sum_{j=0}^{n-2} \overline{B_j} 2^{n+j-1} + B_{n-1}\sum_{i=0}^{n-2} \overline{A_i} 2^{n+i-1}$$

Each AND gate of a multiplier generates partial product bit. The computation steps of a 4 × 4-bit array multiplier is shown in table below. It produces an 8-bit output.

7	6	5	4	3	2	1	0
				a_3	a_2	a_1	a_0
				b_3	b_2	b_1	b_0
				$\overline{a_3 b_0}$	$a_2 b_0$	$a_1 b_0$	$a_0 b_0$
			$\overline{a_3 b_1}$	$a_2 b_1$	$a_1 b_1$	$a_0 b_1$	
		$\overline{a_3 b_2}$	$a_3 b_2$	$a_1 b_2$	$a_0 b_2$		
	$\overline{a_3 b_3}$	$\overline{a_2 b_3}$	$\overline{a_1 b_3}$	$\overline{a_0 b_3}$			

Table

A single AND gate is used to generate each of the product term. Prior inversions are denoted by bars above the variables. Prior to the complete adder or AND gates, inverters are connected to their respective inputs. Product terms are added to each column in accordance with their appropriate weight. $n(n-1) + 3$ complete adders are needed in this system. So, a 4 × 4-bit array multiplier requires 15 adders.

Q21. Draw the Baugh-Wooley with 2's complement generator.

Ans: **Model Paper-3, Q8(b)**

In Baugh-Wooley multiplier, the signed multiplicands are initially converted into their 2's complement forms. Figure below depicts a Baugh-Wooley multiplier with 2's complement generator.

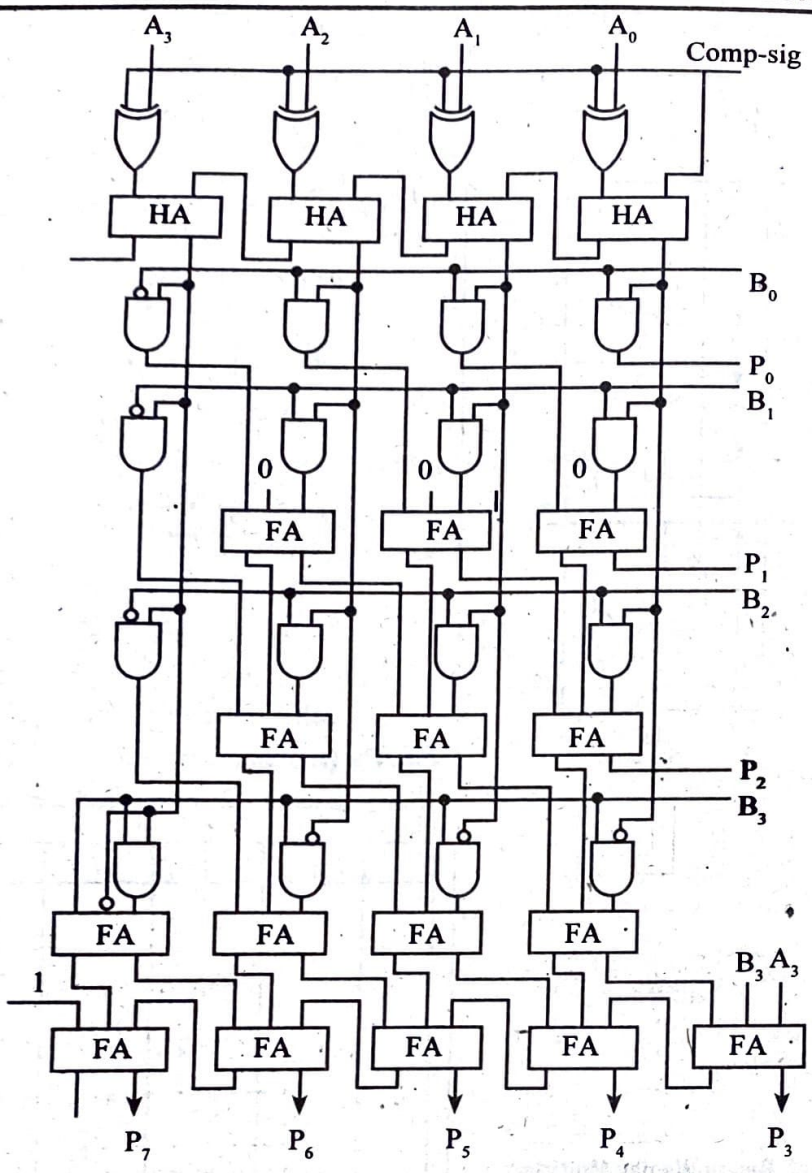


Figure : Baugh-Wooley Multiplier with 2's Complement Generator

It is a combinational circuit that either pass the direct input or in its2's complement form based on control signal.

- If Complementary Signal (Comp-Sig) is high, it converts input to it's 2's complement of form. The XOR gates inverts the input bits and adds 1 to the result. is the output result.
- When the Comp-Sig value is low, input remains same and bit 0 is added to the multiplicand inputs
- After the signed multiplicands are processed, the MSB of the result would reflect the sign of the result (1 for negative, for positive).

Q22. Explain the performance consideration of Baugh-Wooley multiplier.

Ans:

In Baugh-Wooley multiplier, the number of bit operands and the layout techniques affect the area and power consumption of a number of multiplier structures. In a standard-cell-based design flow, increasing regularity and locality at the silicon level reduces power consumption.

Since, the Baugh-Wooley multiplier is an extension of the Braun multiplier, it can use the same optimization interconnect technique used in Braun multiplier.

The modified Baugh-Wooley multiplier with optimized interconnects is depicted in figure below.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

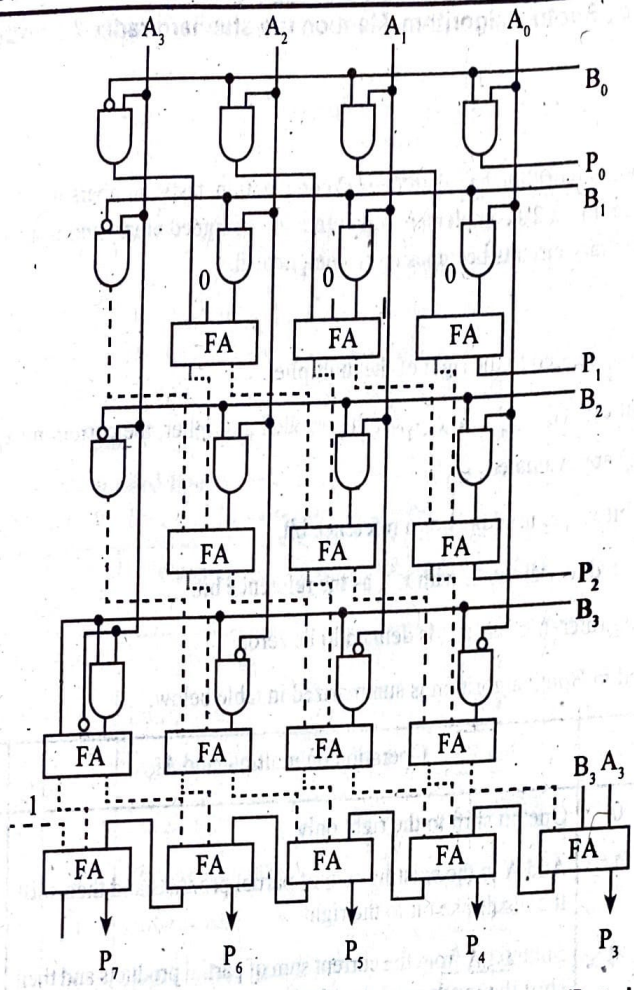


Figure: Modified Baugh-Wooley Multiplier with Optimized Interconnects (Dotted Lines)

5 Booth Multiplier

Points to Remeber
♦ The Booth multiplier uses the Booth encoding algorithm to minimize the number of partial products by evaluating two bits of the multiplier at a time, resulting in a performance gain over conventional multiplier architectures.
♦ This algorithm is valid for both signed and unsigned operands.

3. Explain how Booth's algorithm is suitable for signed number multiplication in comparison of conventional shift and add method.

Model Paper-1, Q9(a)

The Booth's algorithm is an approach that performs multiplication of signed numbers in two's complement form. It is ctional for cases such as, two positive number, one negative number and one positive number or both negative numbers. Unlike current multiplication methods the Booth's method is more or less similar to conventional unsigned multiplication methods with following differences.

Basically, here two bits of multipliers are multiplied inorder to find out which of three courses must occur next. Therefore, algorithm can be specified as,

When the value of the current multiplier bit is set to '1' and the next lower order bit is set to '0', then perform substraction of multiplicand from the partial product.

When the value of the current multiplier bit is set to '0' and the next lower order multiplier bit is set to '1' then perform the addition of multiplicand and partial product.

60

Q24. Explain the operation of Booth's algorithm. Mention the standard radix-2 Booth multiplication rules.

Ans:

Booth's Algorithm

For multiplication, the Booth algorithm, based on radix-2 computation, takes numbers in 2's complement form. Hence is also called radix-2 algorithm. In this, a 2's complement representation of signed binary multiplication is used. As a result, way operand signs are kept in auxiliary circuits becomes more complicated.

Procedure

- Initially an imaginary 0 is appended to the right of the multiplier.
- In order to obtain the i^{th} bit of $y_i (y_{n-1} y_{n-2} \ldots y_1 y_0)$ of the recoded multiplier, the current bit x_i and the previous bit x_{i-1} multiplier $(x_{n-1} x_{i-2} \ldots x_1 x_0)$ are evaluated.
- In this case, the previous bit x_{i-1} is used only as a reference bit.
- Similarly x_{i-1} will be recoded to yield y_{i-1}, with x_{i-2} as the reference bit.
- For $i = 0$, its corresponding reference bit, x_{i-1} is defined to be zero.
- The recoding method used in Booth algorithm is summarized in table below.

x_i	x_{i-1}	Operation on multiplicand A	y_i
0	0	One bit shift to the right only	0
0	1	Add A to the existing sum of partial products and then shift the result one bit to the right	1
1	0	Subtract A from the current sum of partial products and then shift the result one bit to the right	$\bar{1}$
1	1	One bit shift to the right only	0

Table: Recoding Method in Booth Algorithm

Standard Radix-2 Booth Multiplication Rules

- Add a zero to the right of the multiplier's Least Significant Bit (LSB).
- Starting with the low-order bit (LSB) and the added zero, examine groups of two adjacent multiplier bits.
- Shift the partial product 1 bit to the right, if the pair is 00 or 11.
- Multiply the partial product by the multiplicand and shift the partial product to the right one bit, if the pair is 01.
- Subtract the multiplicand from the partial product and shift the resultant partial product to the right by one bit, if the pair is 10.
- Continue with overlapping pairs of bits, with the MSB of one pair becoming the LSB of the next.
- Each iteration through the algorithm eliminates 1 bit of the multiplier number in this manner.
- The partial product is updated according to the rules, except that no shift is made when the last pair of bits is evaluated.

The 2's complement multiplier is recoded considering the fact that a k-long sequence of 1's is the same as a $(k-1)$-long sequence of 0's. This change of a string of 1's to 0's helps to minimize the partial products. Any binary number can be recoded in this way, with a minus 1 (denoted by $\bar{1}$ at the low end, a positive 1 to the left of the high end, and between 0's replacing each group of 1 or more adjacent 1's. This recoding method doesn't have to be done in a specific order. It can even be done at the same time for all bits.

Drawback

Booth's approach becomes inefficient in the presence of isolated 1s.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Q25. Explain modified Booth algorithm. Mention the set of rules used for the radix-4 modified Booth algorithm.

Ans:

In modified Booth algorithm, the multiplier is recoded in the higher radix. Less number of cycles are needed to produce the product because more bits can be inspected and eliminated in every cycle.

Procedure

- A zero is appended to the right of the LSB in the modified Booth's technique.
- With x_{i-2} as a reference bit, the multiplier bits x and x_{i-1} are recoded into y and y_{i-1}.
- In the next step, x_{i-2} and x_{i-3} are recoded into y_{i-2} and y_{i-3}, respectively, with x_{i-2} serving as the reference bit.
- As a result, each multiplier is subdivided into three-bit substrings, with adjacent groups sharing a same bit.
- On the right side of the multiplier x, the term x_{i-1} with a zero value is added.
- The bit structure of MSB becomes $x_1\ x_0(x-1)$, followed by $x_3\ x_2\ (x_1)$, $x_5\ x_4\ (x_3)$ and so on.

Set of rules used for the radix-4 modified Booth Algorithm: The set of rules used for the radix-4 modified Booth algorithm to work for odd number of values i.e., i = 1, 3, 5 ... is shown in table below.

x_i	x_i-1	x_i-2	y_i	y_i-1	Comments	Recoded Digit	Operation on multiplicand A before shifting 2 bits to the right
0	0	0	0	0	String of zeros	0	Nothing
0	0	1	0	1	End of ones	+1	Add A to the existing sum of partial products
0	1	0	0	1	Isolated one	+1	Add A to the existing sum of partial products
0	1	1	1	0	End of ones	+2	Add 2A to the existing sum of partial products
1	0	0	1	0	Beginning of ones	–2	Add –2A to the existing sum of partial products
1	0	1	0	1	Isolated zero	–1	Add –A to the existing sum of partial products
1	1	0	0	1	Beginning of ones	–1	Add –A to the existing sum of partial products
1	1	1	0	0	String of ones	0	Nothing

Table: Modified Booth Recoding for Radix-4 Booth Algorithm

In Booth's recoding method, the carry is not propagated on to the next stage. It is possible to generate each of the recoded digits separately.

Q26. With a neat block diagram, explain modified Booth multiplier in brief. Model Paper-3, Q9(a)

Ans:

The block diagram of a modified Booth multiplier is depicted in figure below.

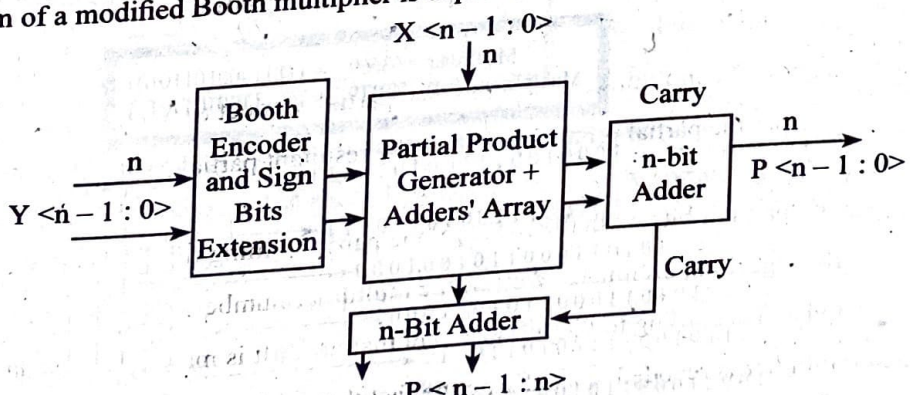


Figure: An n x n Modified Booth Multiplier

1. For concatenating bits, a colon is used. The main components of modified Booth multiplier are,
2. The Booth encoder and sign extension bits
3. An array of multipliers, which consists of partial product generators and 1-bit adders
4. A final stage adder for $2n$-bit addition.

As a result, the Booth encoder generates the following five signals for each of the multiplier's three unique 3-bit segments: 0, $+1A$, $+2A$, $-1A$, and $-2A$.

Q27. Explain the concept of modified Booth multiplier with a suitable example.

Ans:

A radix multiplier of higher radix like radix-4 multiplier/ Booth recording multiplier minimizes the number of adders (and hence the delay needed to generate the partial sums) by analyzing the plurality bits at a time. Here, the multiplier bits are divided into pairs of 2-bits each and at a time 3-bits are scanned. Out of the 3-bits, 2-bits are from present pair and the third bit is higher order bit of adjacent lower order pair. The general procedure of multiplication is shown in figure (1).

Multiplier = 63669_{10} = 1111100010110101
Multiplicand (M) = 40119_{10} = 1001110010110111

```
                                                    M
  1 0 0 1 1 1 0 0 1 0 1 1 0 1 1 1  <-------------  1   Lsb
  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  <-------------  0
                                                    M
  1 0 0 1 1 1 0 0 1 0 1 1 0 1 1 1  <-------------  1   M
                                                    u
  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  <-------------  0   l
                                                    M
  1 0 0 1 1 1 0 0 1 0 1 1 0 1 1 1  <-------------  1   t
                                                    M
  1 0 0 1 1 1 0 0 1 0 1 1 0 1 1 1  <-------------  1   i
                                                    0
  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  <-------------  0   p
                                                    M
  1 0 0 1 1 1 0 0 1 0 1 1 0 1 1 1  <-------------  1   l
                                                    0
  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  <-------------  0   i
                                                    0
  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  <-------------  0   e
                                                    M
  1 0 0 1 1 1 0 0 1 0 1 1 0 1 1 1  <-------------  1   r
                                                    M
  1 0 0 1 1 1 0 0 1 0 1 1 0 1 1 1  <-------------  1
                                                    M
  1 0 0 1 1 1 0 0 1 0 1 1 0 1 1 1  <-------------  1
                                                    M
+ 1 0 0 1 1 1 0 0 1 0 1 1 0 1 1 1  <-------------  1   Msb
-----------------------------------------------------
1 0 0 1 1 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 1 0 1 0 1 1 0 0 0 1 1 = 2554336611_10 = Product
```

Figure (1)

The Booth multiplier has an array of adder cells arranged in rows and the input circuitry is also provided which reduce the multiplier power consumption.

For the generation of all booth recoded control signals, Each recoding cell consists of balanced logic circuitry split bus present in the input circuitry provides first number to the array. The generation of partial products using Booths recoding is shown in figure (2).

Multiplier = 63669_{10} = 1111100010110101
Multiplicand (M) = 40119_{10} = 1001110010110111

```
                                                    0
  1 0 0 0 1 0 0 1 1 1 0 0 1 0 1 1 0 1 1 1  <-- +M   1
                                                    0
                                                    1
                                                    0
  1 1 0 1 0 0 1 1 1 0 0 1 0 1 1 0 1 1 1  <-- +M   1
                                                    0
                                                    1
                                                    1
  1 0 1 0 1 1 0 0 0 1 1 0 1 0 0 1 0 0 0  <-- -M   0
                                                    1
                                                    0
  1 0 1 0 1 1 0 0 0 1 1 0 1 0 0 1 0 0 0  <-- -M   1
                                                    1
                                                    0
  1 1 0 1 0 0 1 1 1 0 0 1 0 1 1 0 1 1 1  <-- +M   1
                                                    0
                                                    0
  1 0 0 1 1 0 0 0 1 1 0 1 0 0 1 0 0 0 1  <-- -2M  0
                                                    0
                                                    1
  1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  <-- -0   1
                                                    1
                                                    1
  0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  <-- -0   1
                                                    1
                                                    0
+ 1 0 0 1 1 1 0 0 1 0 1 1 0 1 1 1        <-- +M   1
                                                    0
                                                    0
-----------------------------------------------------
1 0 0 1 1 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 1 0 1 0 1 1 0 0 0 1 1
```

Figure (2)

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

28. With a neat block diagram, explain the function of Booth encoder. - Model Paper-1, Q9(b)

ns:

The following are the main functions of Booth encoder,

Implements Booth encoding of the three multiplier bits.

Handles the sign extension logic.

A basic structure of Booth encoder is illustrated in figure (1).

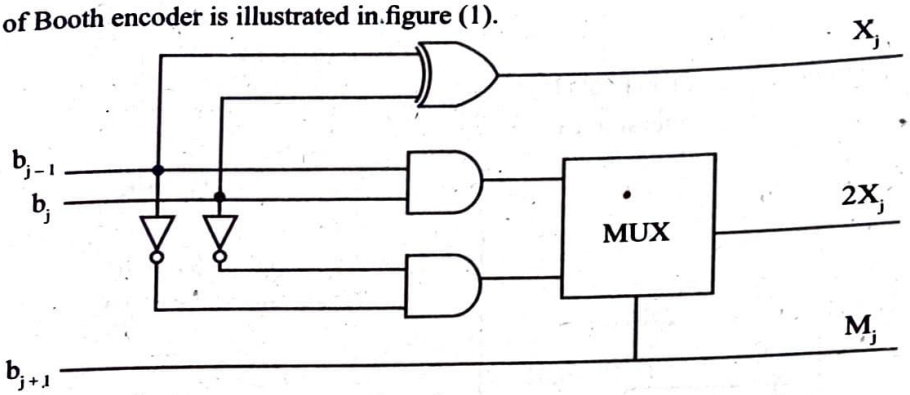


Figure (1): Logic Diagram of a Booth Encoder

The array's encoders are allocated between its various partial products, each of which is assigned to a certain encoder. ly one of the five possible generated partial product signals is high during the steady state operation since there is a circuit for h of the signals. The carry propagation circuits do not have common inputs and do not depend on partial product circuits.

As seen in the block diagram, the signals three signals X_j, $2X_j$ and M_j, are generated from three adjacent bits, b_{j-1}, b_j, and b_{j+1} for selecting a partial product that is 0, $+A$, $-A$, $+2A$, or $-2A$. Here, A is an n-bit multiplicand.

The signals X_j, $2X_j$ indicate whether the partial product is doubled or not and the signal M_j indicate whether negative partial product must be used or not.

An active M_j indicates that negative partial product must be used.

The conventional partial product generator for this implementation is as shown in figure (2).

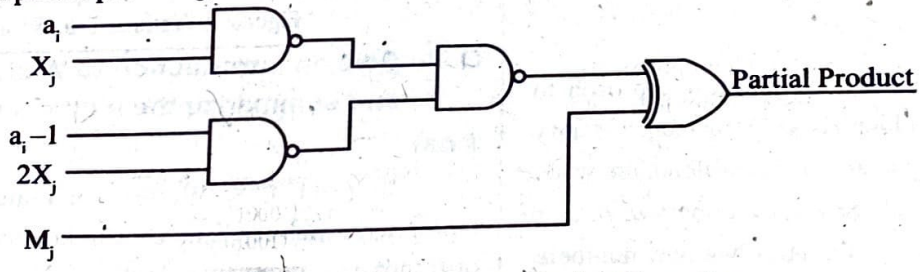


Figure (2): Conventional Partial Product Generator

Using figures (1) and (2), the operations A^* B and $-A^*B$ cannot be directly computed. These can be implemented using a sign-select Booth encoder as shown in figure (3).

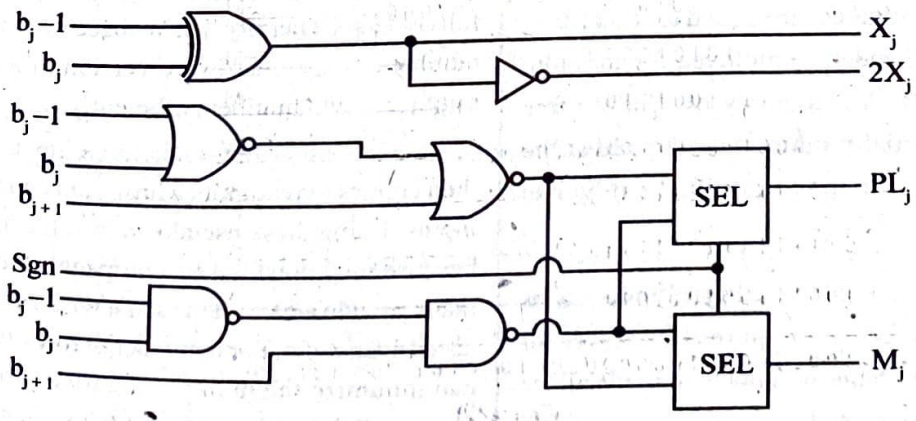


Figure (3): Sign-Select Booth Encoder

In sign-select Booth encoder, the control signal "sgn" performs the signed computations directly.

- The active M_j indicates that negative partial product must be used.
- The active PL_j indicates that positive partial product must be used.

Using sign-select Booth encoder, by replacing the logic circuits with MUXes, the partial product generator can be modified as shown in figure (4).

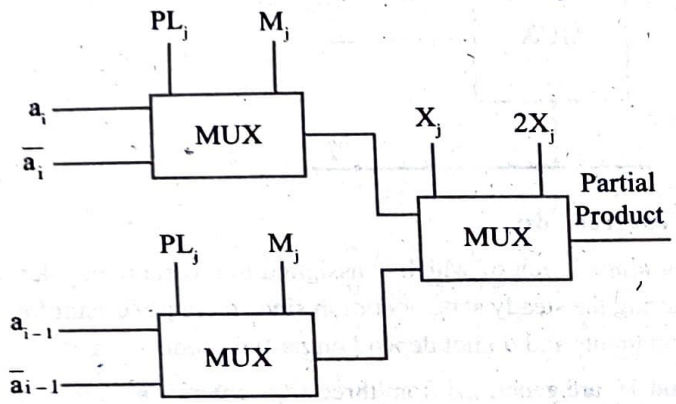


Figure (3): Modified Partial Product Generator

4.6 Introduction to Wallace Tree Multiplier

Q29. Draw the schematic and explain the working of Wallace multiplier.

Ans:

The Wallace tree assembly is the finest approach for speeding up addition. It is an adder tree constructed from carry-save adders. The Wallace tree is an array of full adders whose carry signals are not linked. Three n-bit numbers a, b, c are given by a carry save adder that calculates two new numbers.

Let y and z be two new numbers then $y + z = a + b + c$

The Wallace tree accomplishes 3 to 2 deductions at every single level of the tree, i numbers are united to form $[2i/3]$ sums. In this case a high speed adder is employed for addition, when only two values are left. An assembly of Wallace tree is illustrated in figure. The partial products are initiated at the bottom of the tree. Each of the z outputs is shifted left by one bit because it signifies the carry out.

The greatest advantage of Wallace tree multiplier is it is faster than a simple array multiplier. It requires less power for bit widths between 8 and 32, with the advantage of the Wallace tree growing as word length increased.

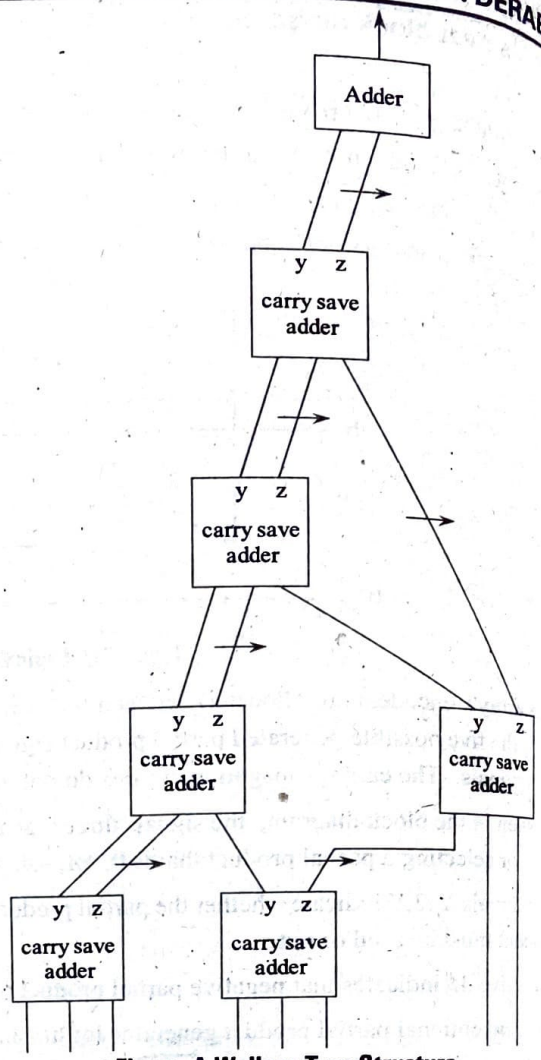


Figure: A Wallace Tree Structure

Q30. Give an introduction to Wallace tree multiplie and summarize the multiplication process.

Ans:

In 1964, C.S. Wallace invented the Wallace Tre multiplier, which can handle the multiplication process for larg operands i.e., more than 16-bit wide numbers. The Wallac tree multiplier uses Wallace tree algorithm. The Wallace tre decreases the number of partial products that must be added in two intermediate results using multiple input compressors an full-adders. Generally, the Wallace tree multiplies two unsigne numbers. It is a hardware version of a digital multiplier tha multiplies two numbers efficiently.

This algorithm adds three inputs using pseudo adders then creates two outputs whose sum equals the sum of the thre inputs. Using these pseudo adders, the multiplication speed u because they don't have to propagate the carry forward. Usin many pseudo adders at the same time can improve multiplicatio speed even more. For n number of rows, the Wallace tree metho can minimize the number of partial products in parallel, wit an overall delay proportional to $\log_{3/2} n$.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

The steps followed in Wallace tree multiplication are,

1. **Partial Product Generation Stage**

In this partial products are generated. These are the intermediate terms that are produced based on the multiplier value. If the multiplier bit is set to '0,' the partial product row is also set to zero, and if it is set to '1,' the multiplicand is copied unchanged. Each partial product row is shifted one unit to the left starting with the 2nd bit multiplication. The sign bit is also extended to the left in signed multiplication.

2. **Partial Product Reduction Stage**

A system of counters reduces the partial product matrix but does not propagate the carries. The obtained matrix consists of the counters' sums and carry. his matrix is then reduced with another set of counters, and the procedure is repeated until a two-row matrix is produced.

3. **Partial Product Addition Stage**

A final adder, generally a carry propagate adder, is used to add up the two rows. To avoid carry propagation until the last adder, this technique uses the carry save architecture. The multiplier's speed is determined by the number of levels in this system.

By using 3:2 compressors, the traditional Wallace tree algorithm decreases the propagation stages.

Q31. Why 4:2 compressors are used in Wallace tree? Explain in detail.

Ans: Model Paper-2, Q9(b)

4:2 compressors were invented by A. Weinberger from IBM in 1981. He proposed replacing the 3:2 compressors in the traditional Wallace tree method with 4:2 compressors to reduce the number of propagation stages in tree multipliers. The advantage of tree multipliers is that their speed scales logarithmically with the length of the operand. Because it can reduce four inputs of the same weight to two, the 4:2 compressor can provide a far more regular structure than the 3:2 counter.

The block diagram and logic diagram of a 4:2 compressor are depicted in figure (1).

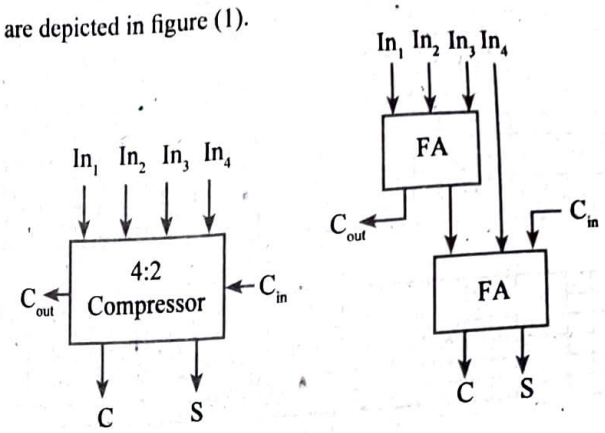


Figure (1): 4:2 Compressor

A 4:2 compressor is a combination of two full adders. It has four input bits In_1, In_2, In_3 and In_4 and two output bits, Sum and Carry (S and C) along with an input carry bit, C_{in} and output carry bit, C_{out}.

The sum, S is expresses as,

$$S = In_1 \oplus In_2 \oplus In_3 \oplus In_4 \oplus C_{in} \quad \ldots (1)$$

Here,

- The input carry bit, C_{in} is the output from the previous lesser significant bit and the output carry, C_{out} is passed on to the next higher significant bit. Both are independent of each other thus avoids carry propagation.
- The Carry, C is associated with the inputs In_4 and C_{in} of first full adder.
- The carry-out, C_{out} is associated with the inputs In_1, In_2 and In_3.

Equation (1) can be viewed as addition of four XOR delays which is equivalent to reducing the partial products output from the Wallace tree structure using 2-layer carry save adders(CSAs).

So, equation (1) can be re-arranged as,

$$S = [(In_1 \oplus In_2) \oplus (In_4 \oplus C_{in})] \oplus C_{in} \quad \ldots (2)$$

With this arrangement, three XOR delays are involved.

Q32. Explain Wallace tree construction. Mention the tree complexities.

Ans:

In general, the Wallace tree multiplies two unsigned numbers. The proposed Wallace tree multiplier architecture consists of an AND array for computing partial products, an adder for adding the partial products so generated, and a sklansky adder at the end of the addition process.

By reducing the number of adders in the partial products reduction stage, the Wallace tree multiplier's latency can be lowered.Multi bit compressors are employed in the proposed architecture to reduce the number of partial product addition steps.

The Wallace tree is constructed by taking all of the bits in each four row at a time and compressing them in the most efficient way possible as shown in figure (1).

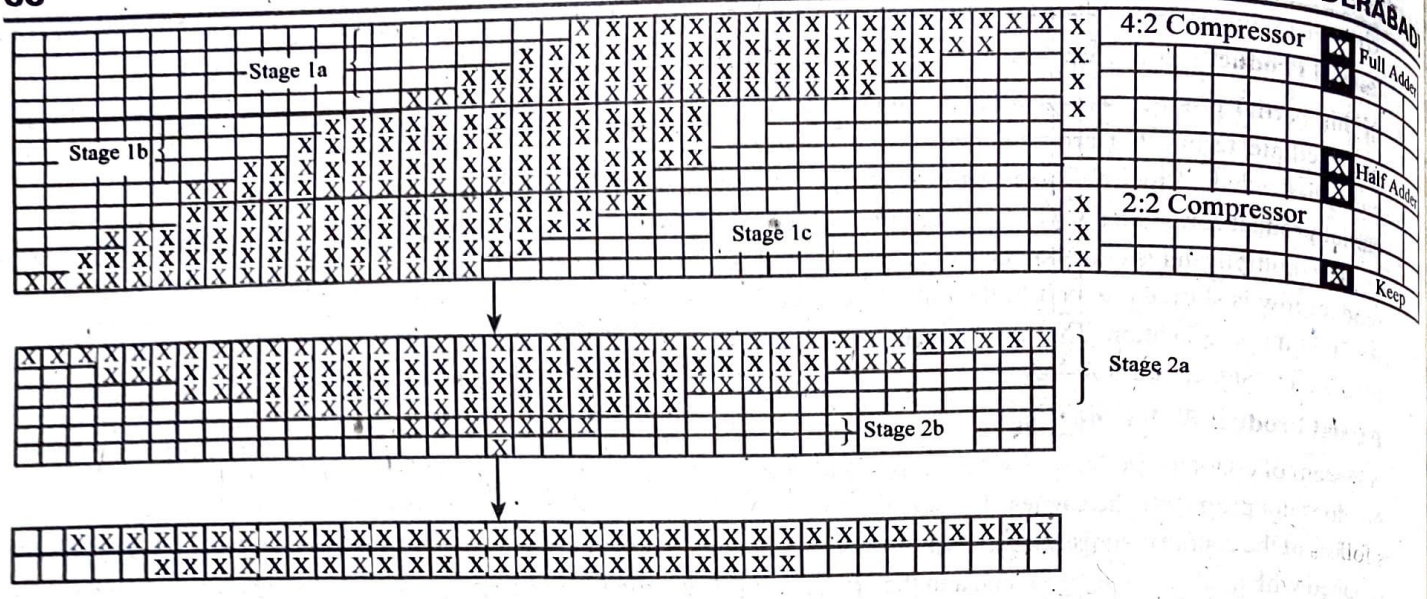


Figure (1): Traditional Wallace Tree Construction

Construction Complexities

1. In traditional Wallace tree, as the partial products are being added in one direction, from to bottom, the number of compressors or adders and wires increases with adding stages. It results increasing the layout width.

   Figure (2) illustrates the increasing width of the adding stages. Each horizontal line in the Wallace tree represents a stage of compressors.

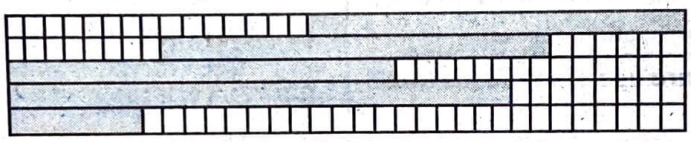


Figure (2): Layout of Wallace Tree Multiplier before Arrangement

2. The compression method varies depending on the number of partial products to be compressed at each position. It's nearly impossible to lay down a Wallace tree multiplier in a rectangular shape without wasting a lot of chip area because of its non-regularity.

   Figure (3) illustrates that when the tree is laid out in a rectangle shape, 38 percent of the overall rectangular area is wasted. The wasted area is called the dead area.

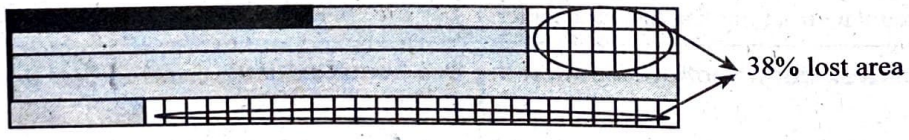


Figure (3): Layout of Conventional Wallace Tree Multiplier after Arrangement

Q33. Explain the modified Wallace tree construction.

Ans: Model Paper-3, Q9(b)

The drawbacks of a traditional Wallace tree can be overcome using the modified Wallace tree. Here, around the centre of the tree, the partial products are divided into two groups, with partial products in one group being added from the top to the bottom and partial products in the other group being added from the bottom to the top.

Figure (1) illustrates the process of splitting a Wallace tree into two parts.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

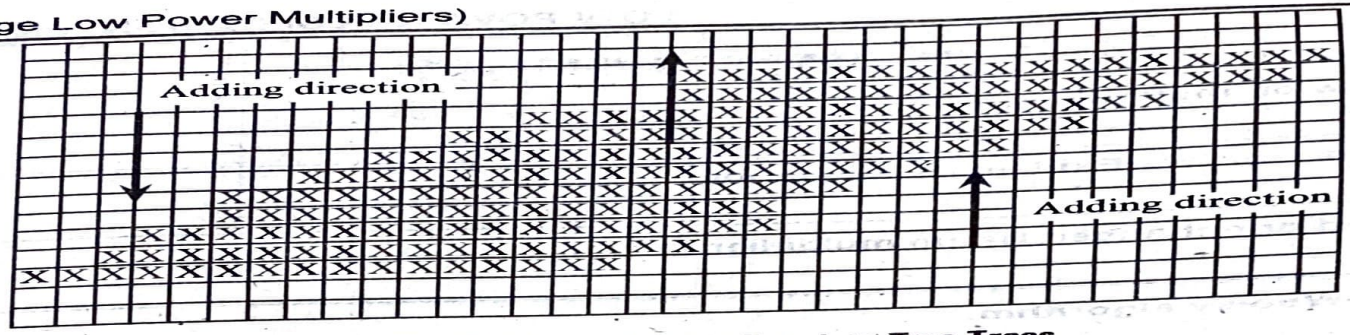


Figure (1): Splitting a Wallace Tree into Two Trees

The steps followed in this process are,

1. The partial product bits are first separated into two trees: the left tree and the right tree. The least significant bits are in the right tree, whereas the most significant bits are in the left tree.
2. The right tree is compressed in a Wallace tree in an upward pattern, whereas the left tree is compressed in a downward pattern. This reduces the amount of wasted area.
3. Then, one of the two sub-trees is accumulates (adds up) downwards, while the other is accumulates (adds up)upwards. A lot of area can be saved as a result.

A comparison of the traditional Wallace tree and its modified version is illustrated in figure (2).

(a) Traditional Wallace Tree **(b) Modified Wallace Tree**

Figure (2): Layout Diagram

IMPORTANT QUESTIONS

Q1. Give an overview of multiplication.

Ans: Refer Q14.

Important Question

Q2. What is Braun multiplier? Explain its operation.

Ans: Refer Q16.

Important Question

Q3. Explain modified or enhanced Braun multiplier.

Ans: Refer Q18.

Important Question

Q4. Explain Baugh-Wooley algorithm.

Ans: Refer Q20.

Important Question

Q5. Draw the Baugh-Wooley with 2's complement generator.

Ans: Refer Q21.

Important Question

Q6. Explain how Booth's algorithm is suitable for signed number multiplication in comparison of conventional shift and add method.

Ans: Refer Q23.

Important Question

Q7. Explain the operation of Booth's algorithm. Mention the standard radix-2 Booth multiplication rules.

Ans: Refer Q24.

Important Question

Q8. Explain modified Booth algorithm. Mention the set of rules used for the radix-4 modified Booth algorithm.

Ans: Refer Q25.

Important Question

Q9. With a neat block diagram, explain modified Booth multiplier in brief.

Ans: Refer Q26.

Important Question

Q10. With a neat block diagram, explain the function of Booth encoder.

Ans: Refer Q28.

Important Question

Q11. Why 4:2 compressors are used in Wallace tree? Explain in detail.

Ans: Refer Q31.

Important Question

Q12. Explain Wallace tree construction. Mention the tree complexities.

Ans: Refer Q32.

Important Question

Q13. Explain the modified Wallace tree construction.

Ans: Refer Q33.

Important Question

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

UNIT 5

Low-Voltage Low-Power Memories

Look for the Logo on the Title Cover Before You Buy
SIA GROUP
SIA GROUP

Syllabus

LOW-VOLTAGE LOW-POWER MEMORIES

Basics of ROM, Low-Power ROM Technology, Future Trend and Development of ROMs, Basics of SRAM, Memory Cell, Precharge and Equalization Circuit, Low-Power SRAM Technologies, Basics of DRAM, Self-Refresh Circuit, Future Trend and Development of DRAM.

Learning Objectives

- Basic architecture of a ROM
- Sources of power dissipation in a ROM
- Future trend and development of EPROM
- Architecture of an SRAM
- Methods to achieve low voltage low power SRAM cell
- Precharge and equalization circuit in an SRAM
- Low power SRAM technologies
- Architecture of a DRAM
- Mechanism of self refresh circuit in DRAM
- Future trend and development of DRAM

Introduction

ROM and RAM are two main memories used in a computer system. ROM is a non-volatile memory that only allows retrieval of prerecorded data.

With is wide range of applications, such as light weight hand-held portable communication systems, semiconductor memory like ROM has become a need in today's computer market. With growing ROM market trends, the demand for low voltage low power ROM increased drastically.

Power requirement reduction in RAM has seen considerable improvement. Many circuit approaches have been invented to provide active and standby power reduction in both static and dynamic RAMs. The growing demand of low power high density electrode applications led to this trend.

PART-A SHORT QUESTIONS WITH SOLUTIONS

Q1. What is a memory?

Ans:

An internal storage area in a computer system which can store the data or information temporarily or permanently is known as 'memory'.

Q2. Interpret Read and Write operation.

Ans:

Read: The process in which stored information is transferred outside the memory is known as 'read' operation.

Write: The process in which new information is stored into memory is known as 'write' operation.

Q3. How does ROM retain information.

Ans:

Model Paper-1, Q1(e)

The abbreviation of ROM is Read Only Memory. ROM is a memory device which stores the binary data permanently. This binary data is written only once during the process of manufacturing and cannot be reprogrammed. This data can't be erased or lost irrespective of power status i.e., whether ON or OFF. Hence, ROM is a non-volatile memory.

Q4. What is the basic difference between the RAM & ROM circuitry?

Ans:

ROM does not have data inputs since it does not perform write operations. However, integrated circuit ROM chips may have one or more enable inputs.

RAM circuit has data input lines besides address selection lines and control lines, since it performs write operation.

Q5. What are the low-power techniques used at architecture level of ROM?

Ans:

The following low-power techniques used at architecture level of ROM.

1. Divided hierarchical word-line structure
2. Selective precharge
3. Difference encoding.

Q6. What are the low-power techniques used at circuit level of ROM?

Ans:

The low-power techniques used at circuit level of ROM are,

1. nMOS precharge
2. Voltage scaling
3. Voltage Keeper
4. Buffer sizing
5. On-chip high-voltage generator.

Q7. Draw the circuit diagram of a loadless 4T SRAM cell.

Ans:

The circuit diagram of a loadless 4T SRAM cell is depicted in figure below.

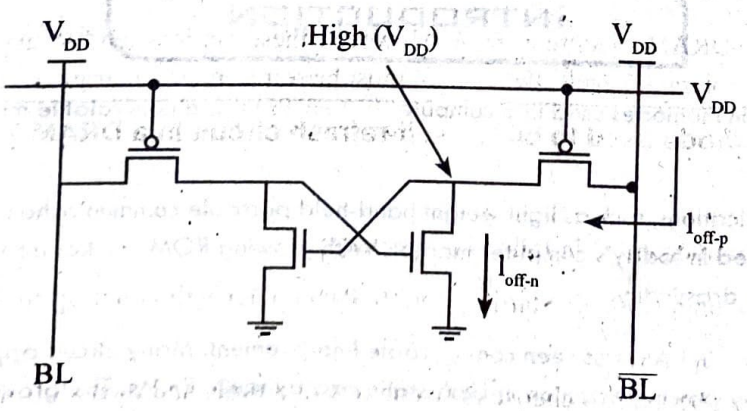


Figure: Loadless 4T SRAM Cell

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Q8. What are the different types of low-power SRAM technologies? Model Paper-2, Q1(j)

Ans:

The different types of low-power SRAM technologies are,

1. Divided word-line architecture
2. AC reduction technique
3. Pulse operation technique
4. Low-power sensing technique
5. Leakage current reduction technique.

Q9. Explain the function of address select control in DRAM. Model Paper-2, Q1(e)

Ans:

- The address logic is used for both row and column addresses in time-slots. So it has two active-low select/control lines $\overline{RAS}$ and $\overline{CAS}$.
- If Row Address select ($\overline{RAS}$) is logic 0 or low, then it latches the row address and begins a memory cycle. The address pin data will determine which row is selected when RAS is enabled.
- If Column Address select ($\overline{CAS}$) is logic 0 or low, then it latches the column address and begins a memory cycle. The address pin data will determine which row is selected when $\overline{CAS}$ is enabled.

Q10. Draw the structure of a 1T DRAM. Model Paper-3, Q1(j)

Ans:

The structure of a 1T DRAM is illustrated in figure below.

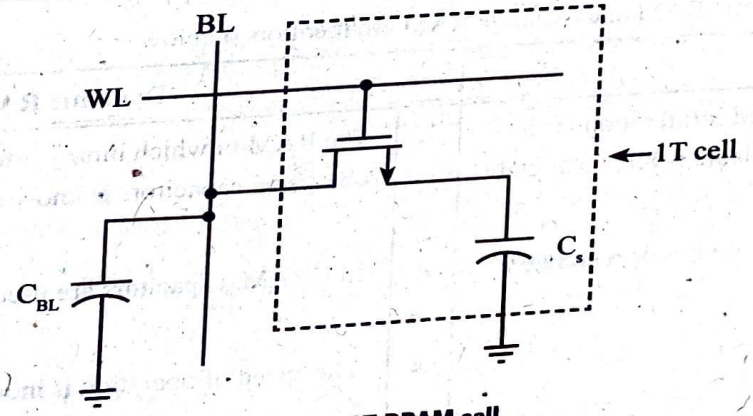


Figure: 1T DRAM cell

Q11. Why a self-refresh circuit is used in a DRAM?

Ans:

The capacitors are used in DRAM to store charges. However, these charges tend to deteriorate over time. As a result, before the stored voltages reach intolerable levels, the DRAM must be refreshed to its original state.

Q12. What are the usual methods used to built a self-refresh circuit in a DRAM ? Model Paper-1, Q1(j)

Ans:

The refresh circuit can be implemented in following two ways,

1. Using an address decoder and a counter: Starting from 0; the counter will count up to the maximum amount of entries possible.
2. Using a shift register: It can be used to choose only one row at a time, and then shift the selection to the next row in a sequential manner. This method requires clock, reset and refresh trigger pins.

Q13. What are the main components of the refresh trigger signal generator?

Ans:

The main components of the refresh trigger signal generator are,

1. Test memory cell
2. Reference signal generator
3. Differential amplifier
4. Differential to digital converter.

Q14. Compare and contrast EEPROM and flash memory.

Ans: Model Paper-3, Q1(

Parameters	EEPROM	Flash Memory
Writable	Possible to write any number of times.	Possible to write any number of times.
Erasable length	One byte at a time.	Block of data at a time.
Speed	Fast to read but it is slow to write and erase data.	Fast to read but it is slow to write and erase data.
Cost per byte	Costly	Moderate

Q15. Compare a static and dynamic RAM cell.

Ans:

The differences between static RAM and dynamic RAM are mentioned below.

	Static RAM		Dynamic RAM
1.	The RAM in which stored information is valid only when power is available is known as static RAM.	1.	The RAM in which binary information is stored as electric charge on capacitors is known as dynamic RAM.
2.	In SRAM, flip flops are used as data storage elements.	2.	In DRAM, capacitors are used as data storage elements.
3.	The speed of operation is high.	3.	The speed of operation is moderate.
4.	It is expensive.	4.	It is inexpensive.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

PART-B ESSAY QUESTIONS WITH SOLUTIONS

5.1 Read Only Memory

5.1.1 Basics of ROM

Q16. Draw and explain the architecture of a basic ROM.

Ans:

The architecture of a basic ROM is depicted in figure (1).

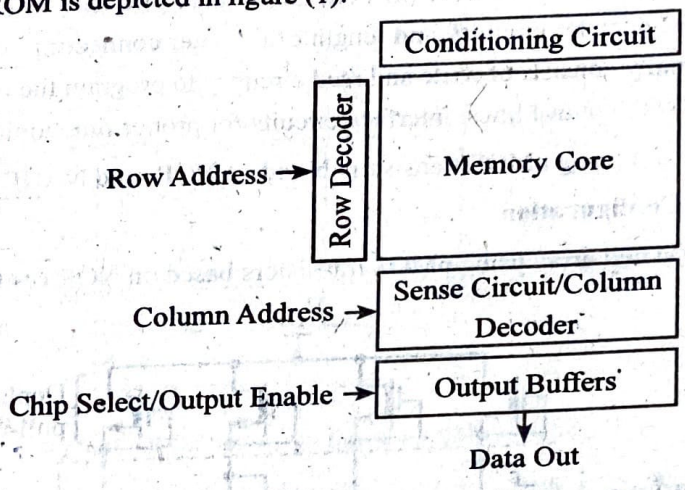


Figure (1): Architecture of a basic ROM

- The row and column address decoders are used to select one of the 2^n words by decoding the n-bit address.
- The information is retrieved or read from the bit-lines using the sense amplifiers, which are precharged by the conditioning circuit.
- Prior to the sensing procedure, column decoders are typically used to choose a set of desired data from the bit lines.
- The output enable signal will then activate the output buffers and transfer the data to the outputs when the data outputs are ready to be accessed.
- The chip select/output enable signal is used to increase the number of words and the size of the words in memory.

The basic ROM consists of a decoder of size $n \times 2^n$ and a memory array. Figure (2) shows a 1 kbit ROM structured as 256 words of 4 bits each.

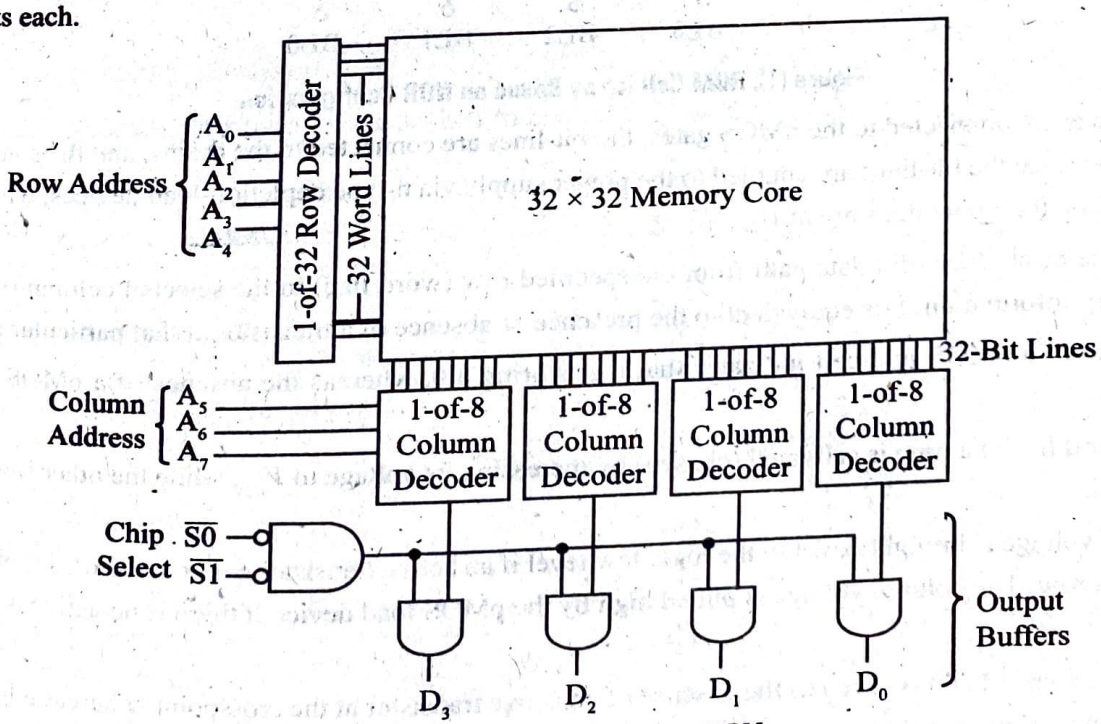


Figure (2): Architecture of a basic ROM

The 8-bit address lines include 5-bit row address and 3-bit column address. A core of memory cells with a 32x32 matrix is formed, having 32 word lines. The row decoder selects one of the 32 rows, while the data selector selects four of the 32 columns from a group of four 8-to-1 column decoders. The selected word-line produces a 4-bit output, namely D0-D3, when combined with the chip choose signal, which feeds the output buffers.

Q17. With neat circuit diagrams, explain the basic structures of ROM cell core using nMOS transistors.

Model Paper-1, Q10(a)

Ans:

The core of the ROM system is the memory array. The storage unit in the memory array is called a cell. It usually stores one bit of data. In some memories, a cell can store more than one bit of data. The array is mostly implemented in square shape so as to reduce the complexity of the decoder circuitry and length of the inter connecting wires. Address decoder identifies the memory cell to be accessed. It internally consists of write and read circuitry to program the memory cells and read out the stored data. The memory chip is provided with control lines, interface circuits for proper functioning.

The ROM cell array can be built using nMOS transistors based on NOR and NAND configurations.

ROM Cell Array Based on NOR Configuration

Figure (1) illustrates the ROM cell array using nMOS transistors based on NOR configuration.

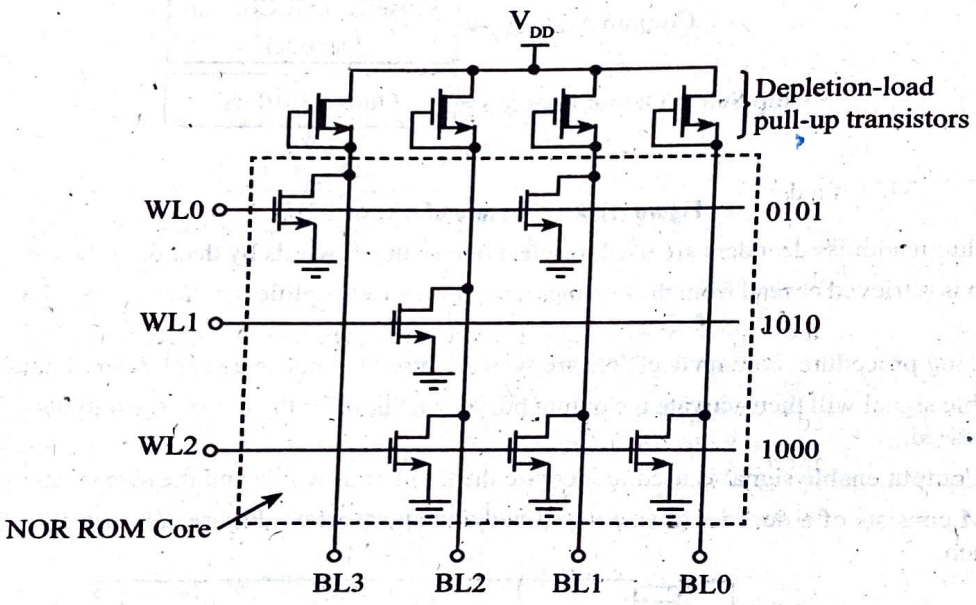


Figure (1): ROM Cell Array Based on NOR Configuration

- The word-lines are connected to the nMOS gates, the bit-lines are connected to the drains, and the sources are connected to ground. Because the bit-lines are coupled to the power supply via n-type depletion-load devices, which serve as conditioning circuits, their potentials are at V_{DD}.
- The presence or absence of a data path from the specified row (word line) to the selected column (bit line) is used for storing binary information. It is equivalent to the presence or absence of a transistor at that particular location.
- The presence of a nMOS in a cell indicates that it is storing a 0, whereas the absence of a nMOS indicates that it is storing a 1.
- Only one word line at a time is activated (chosen) by increasing its voltage to V_{DD}, while the other rows are kept at a low voltage level.
- The column voltage is brought down to the logic low level if an active transistor exists at the cross point of a column and the specified row. The column voltage is pulled high by the pMOS load device if there is no active transistor at the cross point.
- As a result, a logic '1' bit is saved as the absence of an active transistor at the crosspoint, whereas a logic '0' bit is stored as the existence of an active transistor.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

ROM Cell Array Based on NAND Configuration

Figure (2) illustrates the ROM cell array using nMOS transistors based on NAND configuration.

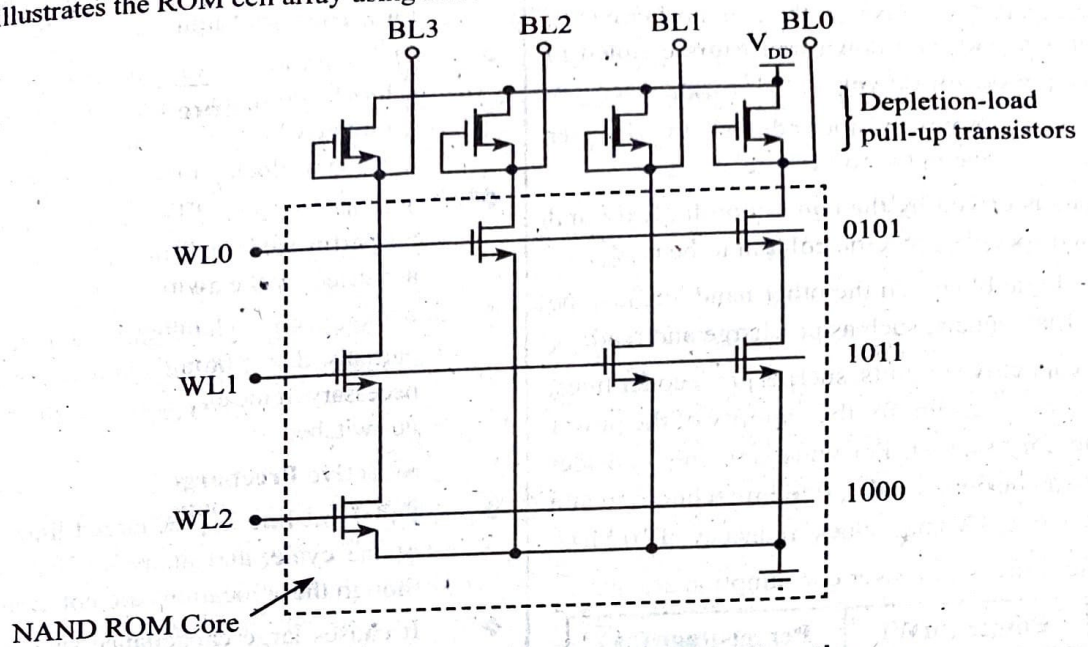


Figure (2): ROM Cell Array Based on NAND Configuration

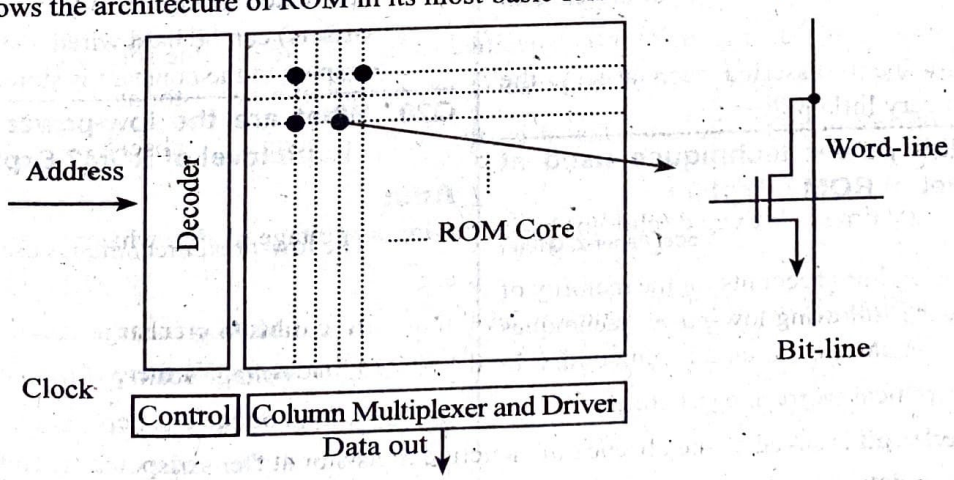
- Each bit line is made up of a depletion-load NAND gate that is driven by some of the word lines' row signals.
- Except for the selected line, which is pulled down to logic-low voltage level in normal operation, other word lines are maintained at logic-high voltage level.
- If a transistor is present at the crosspoint of a column and the selected row, it is turned off, and the load device pulls the column voltage high.
- The column voltage is pulled low by the other nMOS transistors in the multi-input NAND structure if no transistor exists (is shorted) at that particular crosspoint.
- Thus, a logic "1" bit is stored by the presence of a deactivable transistor at the crosspoint, whereas a logic "0" bit is stored by a shorted transistor.

5.1.2 Low-Power ROM Technology

Q18. What are the sources of power dissipation in a ROM? Explain.

Ans:

Figure below shows the architecture of ROM in its most basic form.



Figure

- The memory bank, or ROM core, is made up of an array of transistors designed to store bits of data.
- The NAND and NOR arrays are the two most common types of ROM, with pull-down transistors coupled in series and parallel, respectively.
- Based on the appropriate input address, the decoder selects one word-line in the ROM core at a time.
- The data bus is driven by the column multiplexer and driver, which specify a specific column to be read.
- The control logic block, on the other hand, creates the ROM's internal signals, such as precharge and read.

The highly capacitive circuits, such as predecoder lines, word lines, and bit lines, account for the majority of the power dissipation in a memory system. For understanding, consider a 2 K × 18 ROM was designed using 0.6 pm technology at a power supply voltage of 3.3 V and a clock frequency of 10 MHz.

- In Table, the sources of power consumption are listed.

Block	Power (mW)	Perecentage (%)
Decoder	0.06	2.1
ROM Core	2.24	89
Control	0.18	7.2
Drivers	0.05	1.7

Table

- The precharge of the bit-lines in the ROM core accounts for the majority of the overall power loss. This is owing to the high bit-line capacitance, which is connected to the transistors' drains. In addition, each access switches more than 18 bit lines. This is due to the fact that the word-line picks more bit-lines than are required. If a 12 to 1 multiplexer is used, then, instead of one bit-line, at least four more will switch.
- The control logic consumes a significant amount of power because, it produces the precharge signal for the ROM core and also activates the output drivers and decoding logic.
- As only one word-line is asserted each access, the decoder consumes very little power.

Q19. What are the low-power techniques used at architecture level of ROM? Explain.

Ans: Model Paper-2, Q10(a)

The switching of the bit lines accounts for the majority of the power dissipation. So, the following low-power techniques used at architecture level of ROM.

1. Divided hierarchical word-line structure
2. Selective precharge
3. Difference encoding.

1. Divided Hierarchical Word-line Structure

- The divided word-line approach has been introduced for use in SRAMs.
- The main idea is to split memory into blocks and run a block word line in one layer (metal1 or poly) and a global word line in the other. As a result, just the necessary block's bit cells are accessed.
- The same idea is applied to ROMs. The ROM can be partitioned into multiple blocks, each of which is activated by the address bits.
- Although this technology can minimize power dissipation partially, due to layout considerations, a 4:1 multiplexer is necessary. It means, more than one bit line per bit may still be switched.

2. Selective Precharge

- Every bit line is precharged high during the first half of the cycle, and many bit lines are discharged even though these locations are not accessed.
- It causes large capacitance switching every cycle and results in high power dissipation.
- This problem can be overcome using selective precharge technique in which the bit lines that are only accessed are precharged.
- This approach has a low hardware overhead because most of the control logic is required to drive the multiplexers at the end of the ROM.

3. Difference Encoding

- It is possible to reduce the size of the ROM core by using difference encoding.
- The ROM is sequentially accessed for digital filters and other applications.
- Differences between adjacent data can be stored by ROM cores, rather than the entire value, if there are no significant changes from one address to another.
- The main drawback is that the original value must be calculated using an adder. Different constants (i.e., offsets) can be hard wired and only the difference with respect to the constant is stored.

Q20. What are the low-power techniques used at circuit level of ROM? Explain.

Ans: Model Paper-3, Q10(a)

The low-power techniques used at circuit level of ROM are,

1. nMOS precharge
2. Voltage scaling
3. Voltage Keeper
4. Buffer sizing
5. On-chip high-voltage generator.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

nMOS Precharge

Limiting the voltage swing is a salient technique for minimizing the power dissipation in the bit lines.

This is achieved by using NMOS transistors to precharge bit lines high in the ROM core. Bit lines are precharged to $V_{dd} - V_t$, where V_t is the threshold voltage.

As the bit lines only switch between $V_{dd} - V_t$ and ground, significant power savings are possible.

The disadvantages of this technique are that it degrades noise margins and the body bias effect (which raises the threshold voltage) necessitating careful output driver design.

Voltage Scaling

One of the most effective techniques for decreasing power dissipation is voltage scaling. Voltage scaling makes it simple to get a quadratic improvement. The speed of the circuits degrades as the voltage drops, despite the fact that this technology is quite good at reducing power.

A first order derivation showing the delay of CMOS gates can be expressed as,

$$T_{delay} = \frac{C_L V_{DD}}{I} = \frac{2C_L V_{DD}}{\mu C_{ox}\left(\frac{W}{L}\right)(V_{DD} - V_{th})^2}$$

Where,

C_L – Load capacitance

V_{DD} – Power Supply voltage

μ – Electron mobility

C_{ox} – Oxide capacitance

W – width of the channel

L – Length of the channel

V_{th} – Threshold voltage

Since the transistor driving the bit lines is close to minimum, the speed of ROMs is greatly reduced.

Voltage Keeper

Switching of bit lines is minimized once the number of non-zero terms has been reduced.

Even if the same ROM address is accessed multiple times, bit lines must be precharged each time.

A voltage keeper is used to decrease switching in the data bus or the adder that is required to convert from sign magnitude to two's complement.

Figure (1) depicts a possible keeper implementation using invert logic.

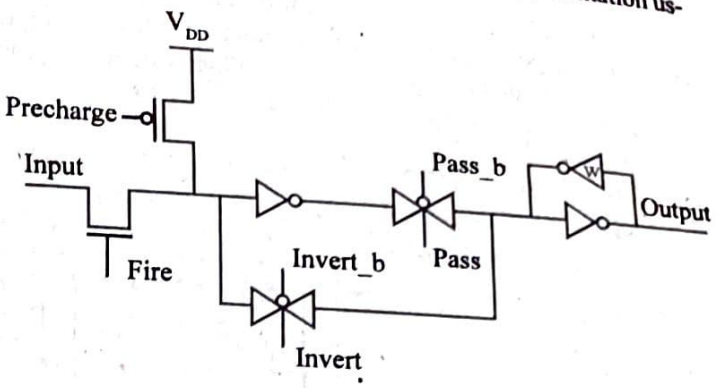


Figure (1)

- The voltage keeper stores previous data and prevents transitions in the data bus and adder (if sign magnitude is used). After the ROM core has been assessed, the fire signal is enabled.
- If sign magnitude or Row Invert are performed, pass and invert signals are used.

4. Buffer Sizing

A large set of buffers is required in the control logic to drive the address lines through the decoder, generate the control signals for the column multiplexers, drive the row lines and drive the precharge signals.

- Figure (2) depicts the model used.

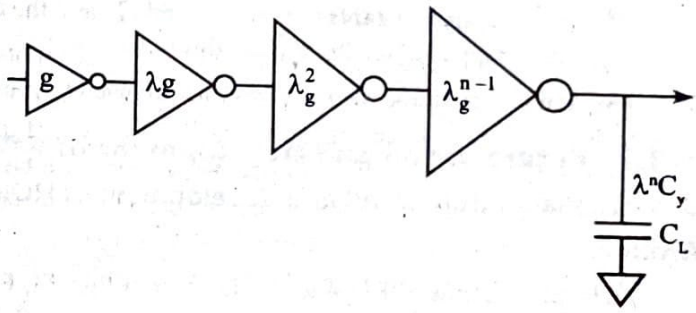


Figure (2): Driving Large Capacitive Loads

- In the figure g represents the conductance while λ represents the taper defined as:

  $(W/L)_{k+1} = \lambda(W/L)_k$

  Where,

  W – width of the transistor

  L – Length of the transistor
- In this case λ indicates the size of stage $k+1$ relative to stage k.
- The number of stages required for a given capacitive load is given by,

  $$n = \frac{\ln C_L/C_i}{\ln \lambda}$$
- This model doesn't take into account parasitic capacitances at the output of each stage.

5. On-chip High-Voltage Generator

A radio frequency identification electrically erasable programmable read-only memory (RFID–EEPROM) requires a high-voltage generator (HVG) to generate a controlled output voltage that is higher than the power source voltage. An on-chip high-voltage generator circuit for low-voltage EEPROMs composed of a pMOSFET-based charge pump circuit driven by boot-strapped clock generators is illustrated in figure (3).

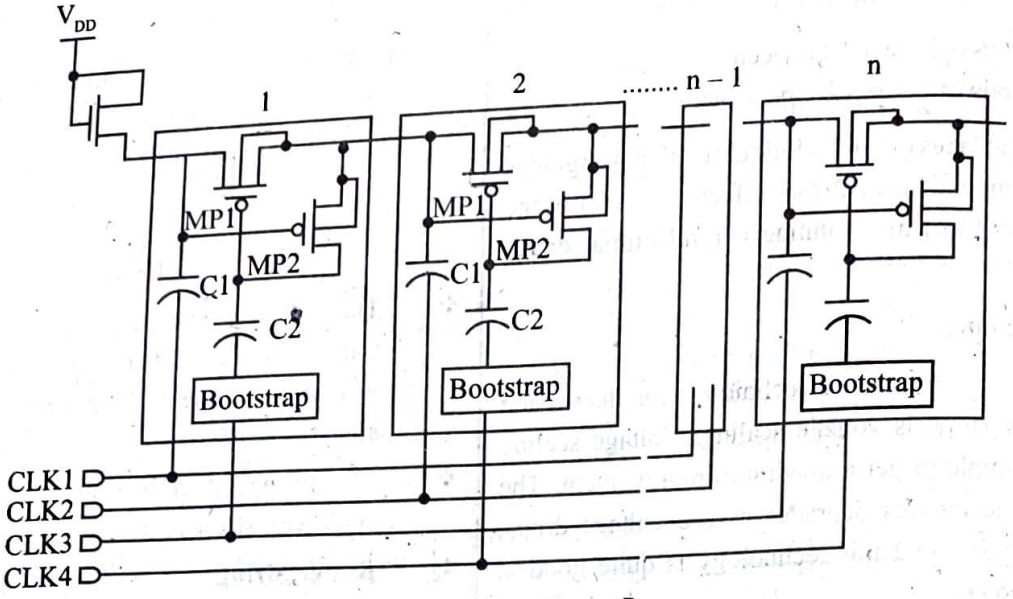


Figure (3): Low-voltage Charge Pump

- The voltage gain per unit stage does not suffer from the threshold voltage drop. The device implemented in a 1.2 /spl mu/m CMOS technology operates as low as 1 V.
- A drop in the threshold voltage has no effect on the voltage gain at every unit stage. This charge pump is capable of generating sufficient programming voltage with a power supply as low as 1 V.
- The clock-coupling capacitors are C_1 and C_3, and the charge-transfer transistors are M_1 and M_3, whose gates are driven by the bootstrapped clock generators through C_2 and C_4.
- M_2 and M_4 are transistor switches that are used to pre-charge the gates of M_1 and M_3.

5.1.3 Future Trend and Development of ROMs

Q21. Explain future trend and development of ROM.

Ans:

The main trends of ROM technology are EPROM, EEPROM, and flash memory.

Table (1) compares the characteristics and distinctiveness of EPROM, EEPROM, and flash memory that influence their commercial viability in the memory market.

	Functionality			Cost			
	Program	Field Erase	System Erase	Type of Package	Low Cost Package	High Density	Low Cost Test
EPROM	Yes	Yes	No	Ceramic with window	No	Yes	No
EPROM	Yes	Yes	Yes	Plastic	Yes	No	Yes
Flash Memory	Yes	Yes	Yes	Plastic	Yes	Yes	Yes

Table (1): Comparison of EPROM and Flash Memory

Development of EPROM

- Despite its high package cost and relative inflexibility, the UV-EPROM has developed a considerable portion of the non-volatile data storage industry due to its high-density field alterable characteristics.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Development of EPROM

- Because of EEPROM's system adaptability, it has carved out significant market niches in the low-density domain below 256 K.
- Semiconductor technology is constantly advancing at a breakneck pace. Over the previous two decades, process technology has seen a considerable drop in size of 12% every year. To keep up with the ever-increasing density of circuits and number of functionalities per chip, new technologies for various semiconductor applications, including memory implementation, have emerged.
- STMicroelectronics' Flash+ idea, for example, combines flash memory with EEPROM on the same chip. Because flash memory enables for dynamic software modification, it is the best nonvolatile memory for storing microcomputer codes.
- EEPROM is the best choice for storing parametric data because it is byte-alterable. As a result, Flash+ memory combines the advantages of flash memory with EEPROM.

Development of Flash Memory

- Flash memories are becoming more popular as a result of their high density, low-cost process and package, and high system flexibility. They have successfully emerged to become the primary centre of high demand dominating the DRAM. This is because, of all semiconductor memory architectures, flash memory has the lowest cost per bit.
- When compared to DRAM, its cell size is substantially smaller, and it requires far less processing steps.
- Wireless consumer items such as cellular phones, digital still cameras, internet audio devices, handheld computers, and set-top boxes are the most powerful drivers of flash memory technology development.

Table (2) shows the technology roadmap.

Year	1998	1999	2000	2001	2002/3
Process	0.35 μm	0.25 μm	0.18 μm	0.15 μm	0.13 μm
Products	16 Mb	32 Mb	64 Mb	128 Mb	265 Mb-1 Gb

Table (2)

By storing more than one data bit into a memory cell, memory suppliers hope to increase the density of flash memory (its capacitor). Newer flash memory, assigns four or more possible quantities of charge per cell, allowing each memory cell to store two or more bits. This initiative will further reduce the cost per bit, making flash memory a more desirable choice.

5.2 Static Random Access Memories

5.2.1 Basics of SRAM

Q22. With neat block diagram, explain the architecture of a static RAM(SRAM). **Model Paper-1, Q10(b)**

Ans:

The block diagram of a Static RAM(SRAM) architecture is depicted in figure (1).

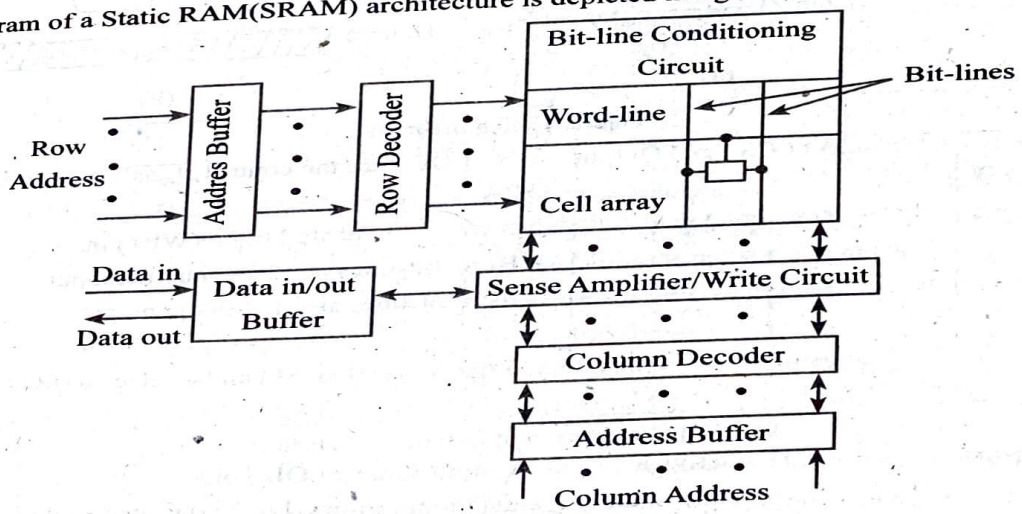


Figure (1): Block Diagram of a Static RAM Architecture

The major blocks of the SRAM are,

1. Address buffers
2. Row and column decoders
3. The memory array
4. Sense amplifiers
5. Input/output buffers.

The precharging of the bit lines is activated by the conditioning circuit. Cells on the memory chip array are either already storing or going to store the data bits (e.g., as a read and write operations). A word line and a pair of bit lines connect each cell. Each R/W operation selects a certain cell by activating its word- and bit-lines.

As a result, the address buffers drive the row decoder and the column decoder, allowing for cell selection. At the point where the activated word line and the activated bit line join, the addressed cell is identified. Then, a column sensing amplifier uses the memory complementary bit-lines to detect the contents of the selected cell in the form of voltage fluctuations.

Critical path from address input to data output determines access time as indicated in figure (2).

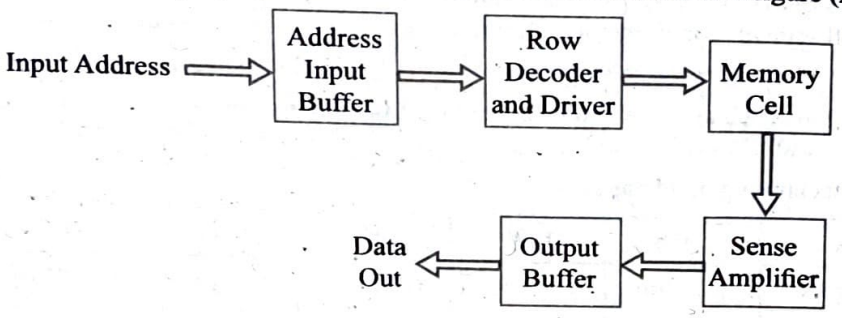


Figure (2)

Q23. Draw the timing diagram of a static RAM (SRAM).

Ans:

Figure below shows the timing diagram of read and write operations of an SRAM.

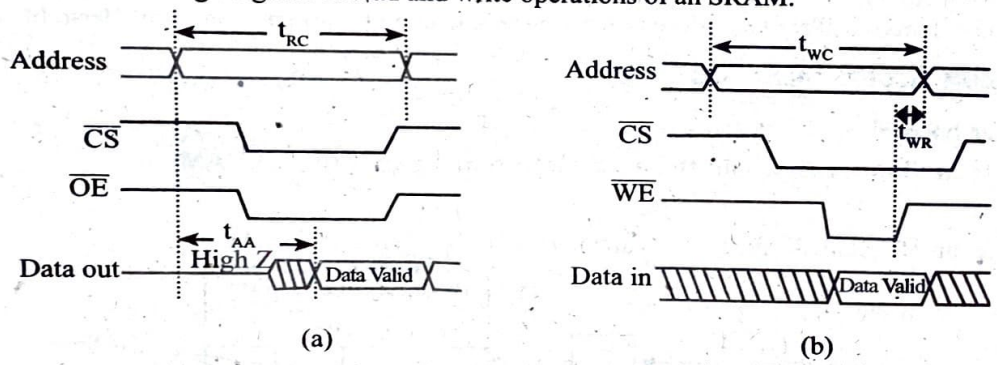


Figure: Timing Diagram

Write Enable ($\overline{WE}$), Chip-select ($\overline{CS}$) and Output Enable ($\overline{OE}$) are the control signals used in SRAM.

1. **Write Enable ($\overline{WE}$)**

- Read and write operations are performed using an active low input Write Enable(WE) pin.
- When WE= 0 or logic low, data is written into SRAM. Here, bit-lines are driven by the input.
- When WE= 1 or logic high, data is read from SRAM. Here, bit-lines are sent to the output.

2. **Chip-select ($\overline{CS}$)**

- In a stack of several SRAMs inside a memory system, a particular SRAM can be selected from the memory bank using an active low Chip-select ($\overline{CS}$) pin.

3. **Output Enable ($\overline{OE}$)**

- Data buffering is carried out using an active low input Output Enable ($\overline{OE}$) pin.
- When $\overline{OE}$ = 0 or logic low, it allows data to enter the data input/output buffer during a read operation.
- SRAM During a write process, $\overline{OE}$ is deactivated.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Read Operation

In read operation, the data stored in SRAM is read out. The read cycle timing parameters are,

1. Read cycle time, t_{RC}
2. Address access time, t_{AA}

Write Operation

In write operation, the data is written into SRAM. The write cycle timing parameters are,

1. Write cycle time, t_{WC}
2. Write recovery time, t_{WR}

5.2.2 Memory Cell

Q24. What is a memory cell? Explain the basic 6T SRAM cell consist of cross coupled CMOS inverters.

Ans:

Memory Cell

The memory cell is the basic structural building unit of computer memory. The memory cell stores one bit of binary data. It must be set to store logic 1 and should be reset to store logic 0. The one-bit data is preserved (stored) until the occurrence of the next set/reset operation. One can access the memory cell to read the data stored in it. The size of memory cell is very crucial in the design of low power high density SRAMs as they are indivisible part of modern VLSI systems.

6T SRAM cell

The standard SRAM cell containing six Transistors is the most basic SRAM cell. Figure (1) illustrates the schematic of standard six transistor SRAM cell. These cell contains two pass transistors (MN3 and MN4) and two cross-coupled CMOS inverters (MP1, MN1and MP2, MN2) with a common read and write port. The two cross-coupled inverters act as flip-flop. With the help of pass transistors MN3 and MN4, the data can be either accessed or written into the cell. These two cross coupled inverters are used for storing one bit of information at a time (either 0 or 1). The Word-Line signal WL controls the pass-transistors (MN3 and MN4) that are joined to the two complimentary bit-lines BL and . They function as transmission gates, allowing bidirectional communication between the flip-flop and the bit lines BL and $\overline{BL}$.

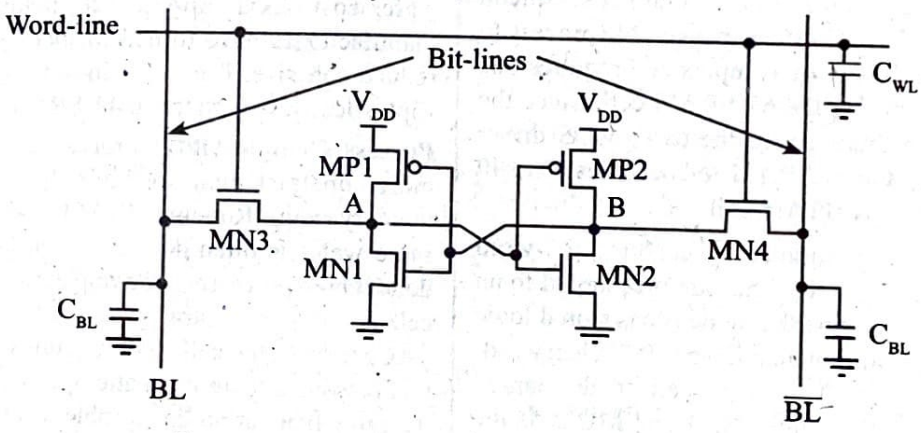


Figure (1): Basic Six Transistor SRAM Memory Cell

The voltages on both bit-lines are precharged to an equalized potential before the read operation. If we assume that node has stored logic "1," then node B stores a complimentary logic "0." When word-line WL is asserted to select this memory cell, discharged to the ground through the transistors MN4 and MN2. At the same time, Current will flow from V_{DD} to BL using the ansistors MP1 and MN3. As a result, the bit-lines show a modest potential difference. The voltage level at BL is slightly greater an in this scenario.

In the case of write operation, the data bit that is intended to be written is sent to BL, while the complement data bit is sent to be written. In other words, if we intend to write "1", then BL is set (raised) to "1" and is reset (lowered) to "0." When the ord-line WL selects the cell, the pass transistor MN3 writes a "1" at node A and the pass transistor MN4 writes a "0" at node The high gain of the flip-flop causes nodes A and B to flip to opposing voltages. Unless there is another write operation, this te, which is a stored 1, will stay the same.

Q25. Explain the 4T SRAM memory cell.

Ans: Model Paper-2, Q10(b)

Even though the six transistor SRAM cell is the most basic standard SRAM cell, it consumes space than the DRAM, which has only one transistor and one capacitor. For that reason, to accomplish a reduction in space used by 6T SRAM cell, a relatively simple design, termed, four transistor SRAM cell was developed. Figure (1) illustrates the schematic of four transistor SRAM cell. The 4T SRAM contains four NMOS and two poly-load resistors. The PMOS transistors of the 6T cell are replaced by very high polysilicon resistors to reduce transistor count and area consumed by the cell. The transistors MD1 and MD2 act as driver transistors, while the transistors MA1 and MA2 act as pass transistors.

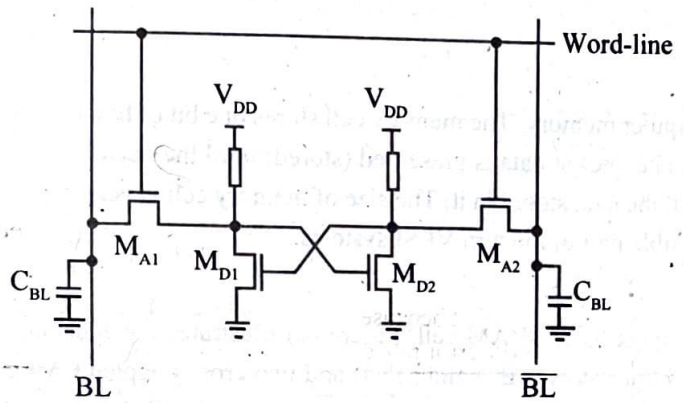


Figure (1): Four Transistor SRAM Memory Cell

One of the most critical requirements for realizing high density memory chips is low power consumption. The power dissipation of this memory cell is reduced to a minimal level by constituting the resistor in an extra high-ohmic polysilicon layer that yields large resistive loads. However, the process to develop the 4T SRAM cell is more complex and requires one additional fabrication step over the 6T SRAM cell. Since the two polysilicon loads are built above the two nMOS driver devices (MD1 and MD2), the 4T SRAM cell occupies 30 to 40 percent less space than the 6T SRAM cell.

The read operation is initiated by enabling the word line (WL) and voltages on both bit-lines are precharged to an equalized potential. If we assume that node A has stored logic "1," then node B stores a complimentary logic "0." When word-line WL is asserted to select this memory cell, is discharged to the ground through the transistors MA2 and MD2. At the same time, Current will flow from VDD to BL through the polysilicon load and the transistor MA1. As a result, the bit-lines show a modest potential difference. The voltage level at BL is slightly greater than in this scenario.

In the case of write operation, the data bit that is intended to be written is sent to BL, while the complement data bit is sent to to be written. In other words, if we intend to write "1", then BL is set (raised) to "1" and is reset (lowered) to "0." When the word-line WL selects the cell, the pass transistor MA1 writes a "1" at node A and the pass transistor MA2 writes a "0" at node B. The high value of loads causes nodes A and B to flip to opposing voltages. Unless there is another write operation, this state, which is a stored 1, will stay the same.

Q26. Explain the racing between 6T SRAM and 4T SRAM memory cells.

Ans: Model Paper-3, Q10(b)

Since its introduction in the 1970s, the 4T SRAM cell has dominated the independent SRAM market due to its smaller area than its 6T equivalent. At the same time, the 6T SRAM cell has remained a popular on-chip storage option. In spite of its bigger size, the 6T SRAM cell started to conquer the 4T SRAM cell in the early 1990s and began to rule the independent SRAM market. The reason for this tendency stems from the benefits provided by the advanced 6T SRAM cell. In order to achieve reduced cell space, the standard 6T SRAM cell used advanced process steps such as Self-Aligned Contacts (SAC) or Local Interconnects (LI) and became an advanced 6T SRAM cell (A-6T SRAM).

The SRAM cells are divided into three categories: 4T cells, Simple 6T (S-6T) cells, and Advanced 6T (A-6T) cells. A major concern for all of these cells is their process complexity, process compatibility, memory cell stability at low voltage, and soft-error rate.

Process Complexity: By incorporating additional processing steps, the smallest memory cell can be developed for a given gate length. In other words, an increase in process complexity decreases the area or size proportionally. But in most cases, an increase in process complexity always contributes to an increase in wafer cost and a lower number of good dice. The advancement in area shrinking overcomes the drawbacks of reduced good die production and increased process complexity. As a result, these tradeoffs will still lower production costs. If we produce a greater number of good dies, then process complexity is not an issue. It will largely compensate for the wafer cost. As a consequence, many of the leading SRAM manufacturers have turned to increasing production steps to reduce cell size. Processes in advanced 6T cells can be less sophisticated than those in 4T SRAM cells.

Process Compatibility: Process compatibility is the second most important thing to think about while designing SRAM memory cells. Running SRAM and logic operations on the same wafer is often advantageous, however this can only be done if both processes are compatible. The previously described cell size reduction strategy via SRAM process upgrades (e.g. SAC) reduces the cell's compatibility with logic processes. For this reason, if there is an attempt to build SRAM and logic on the same fabrication line, problems in production will arise due to an incompatibility among the two. This incompatibility is something that the A-6T SRAM cells will address.

Memory Cell Stability: The stability of the memory cells is the third area of concern, and it is determined by the static noise margin (SNM). The SNM comparison of 4T and 6T cells at three distinct cell ratios is illustrated in Figure (1). [In the figure, β is the ratio of the gate width of the NMOS storage transistor to that of NMOS pass transistor.] When the VDD is lower, the 6T cells are significantly more stable compared to the 4T cells. However, because 4T cells are less stable when exposed to low VDD, they are supposed to have elevated cell ratios than 6T cells. This effect distorts the cell size advantage offered by the traditionally processed 4T cells.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

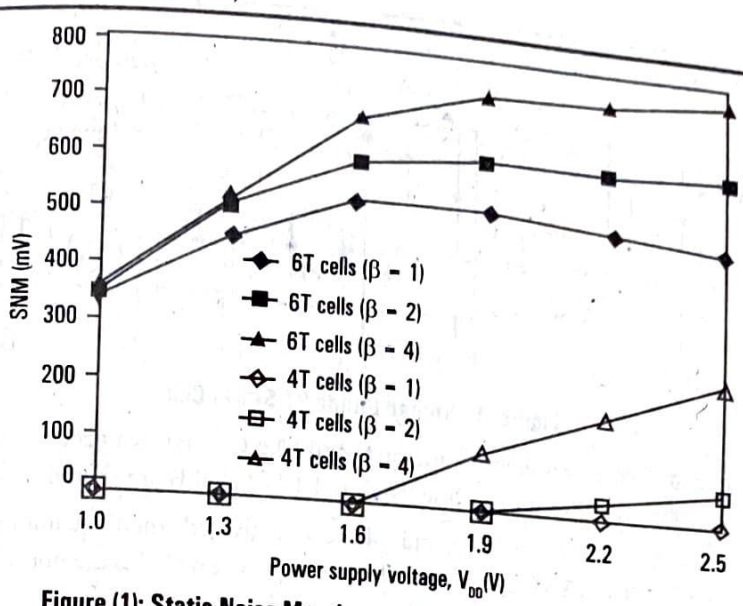


Figure (1): Static Noise Margin as a function of power supply V_{DD}

Soft-Error Rate

The Soft-Error Rate (SER) is the fourth area of concern. In terms of soft error rate, the 6T SRAM cells are stronger than the 4T SRAM cells. This is because 6T cells are larger, and they store a greater amount of charge. The SER becomes a very important issue, when the size and voltages of full 6T SRAM cells are scaled down and lowered a lot.

The advanced 6T SRAM cells started to take over the independent SRAM market because A-6T SRAM has superior process complexity, process compatibility, and memory cell stability at low voltage, and soft-error rate compared to 4T SRAM cells.

Q27. Distinguish between 6T SRAM cell and 4T SRAM cell.

Ans:

The difference between 6T SRAM cell and 4T SRAM cell is as follows,

	6T SRAM Cell		4T SRAM Cell
1.	6T SRAM cells which use six bulk transistors. 6T SRAM contains four NMOS and two PMOS transistors.	1.	4T SRAM cells which use four bulk transistors. 4T SRAM contains four NMOS transistors and two poly-load resistors.
2.	This cell is simple in design and consumes less area and thus can be used to design high-density memory chips.	2.	This cell comprises of lesser number of transistors as well as consumes lesser area compared to the 6T cell. 4T SRAM cell occupies 30 to 40 percent less space than the 6T SRAM cell
3.	Power consumption of 6T SRAM is more compared to 4T SRAM.	3.	Power consumption of 4T SRAM is less compared to 6T SRAM.
4.	The process to develop the 6T SRAM cell is less complex over the 4T SRAM cell.	4.	The process to develop the 4T SRAM cell is more complex and requires one additional fabrication step over the 6T SRAM cell.
5.	This cell fails to maintain its stability, when the technology is scaled down. It suffers from read and write instability, has high power consumption and increased access time when voltage scaling is done.	5.	This cell is sensitive to noise and soft error because of the very high resistances involved in the design.

. With neat diagram, explain any one design method to achieve low-voltage low-power SRAM cell.

:

SRAM cells are accessed often and have high capacitance buses. So, they dissipate power. A low-voltage low-power -P) SRAM cell can be designed that can operate well even at low supply voltage. One of the design method is shared bit-line M configuration.

To achieve ultralow power dissipation in SRAM macrocells, the shared bit-line SRAM cell with a modified address as-ent is depicted in figure (1). To ensure high read sensitivity, a MUX-merged charge-transfer sense amplifier is used.

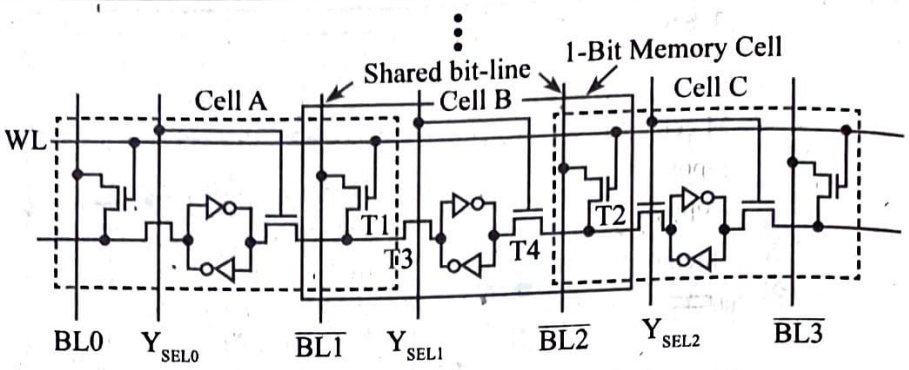


Figure (1): Shared-bitline 8T SRAM Cell

- In this configuration, to a standard six transistor memory cell (i.e Cell B) two access transistors, T1 and T2, are added. $\overline{BL1}$ and T1 of Cell B are shared with Cell A while BL2 and T2 of Cell B are shared with Cell C.
- When WL and YSEL1 select Cell B, T1, T2, T3, and T4 are all activated, and the datum stored in Cell B is accessible to $\overline{BL1}$ and BL2. Because YSEL0 and YSEL2 are not selected, Cell A and Cell B are not selected. As a result, no memory cell current is wasted.
- T3 and T4 are turned off when Cell A is selected, and the datum stored in Cell A is accessible to BL0 and $\overline{BL1}$.

Advantages

- Because both cells are picked at different time intervals, data in a bitline shared by two adjoining cells does not conflict.
- This architecture decreases the number of transistors in a memory cell to 7 equivalently, reducing the area penalty while avoiding memory-cell current waste.
- To prevent high parasitic capacitance to a bitline, a Y_{SEL1} controls T3 and T4, which are not directly linked to bitlines. Furthermore, because the number of transistors connected to a wordline is reduced by nearly half, the parasitic capacitance of wordlines is reduced. This improves wordline selection speed while lowering power consumption due to charge and discharge.

Because two transistors, such as T3 and T4, are coupled to the Y_{SEL} in each memory cell, the parasitic capacitance of the Y_{SEL} is higher than that of a bitline. Charge and discharge power dissipation to this parasitic capacitance should be minimized. Because the Y_{SEL} is controlled by a column address, the column address assignment can be modified. That is, more significant bits in a memory address are assigned to the column address. Because the change probability of a more significant bit is lower than that of a less significant bit, this updated address assignment allows the Y_{SEL}'s effective operating frequency to be reduced and the power consumed by the Y_{SEL} to be reduced. As a result, the new SRAM cell layout eliminates wasted memory-cell current while reducing the area penalty.

Peripheral Circuit Technique for Shared-Bitline SRAM Cell Architecture

Since two access transistors, such as T1 and T3 or T2 and T4 in figure (1) are coupled serially, the memory-cell current is minimized. The bitline signal is reduced, and access time is increased. To avoid access time from rising, the MUX-merged charge-transfer sense amplifier was designed as shown in figure (2).

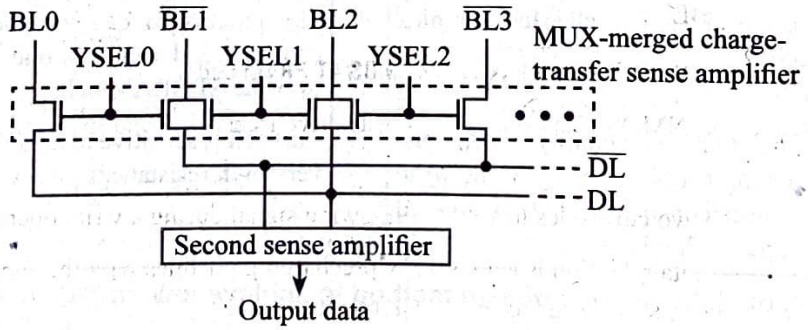


Figure (2)

The decreased bitline signal is amplified by a charge-transfer amplifier in a selected transistor in the MUX. The output signal of the MUX-merged sense amplifier is then amplified by the second sense amplifier to CMOS level signal. Due to the lower memory-cell current, this reading circuit method improves the sensitivity of the read operation while reducing the rise in access time.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Q29. Explain power-efficient 7T SRAM cell with current-mode read and write configuration.

Ans:

The power-efficient 7T SRAM cell with current-mode read and write configuration is one of the design techniques for high performance, low-voltage low-power(LVLP) SRAM cell. It is preferable to use a 7T memory cell that uses the current-mode scheme for both read and write operations since the voltage fluctuations at bit-lines and data-lines are kept to a minimum.

The circuit schematic of a 7T memory cell is depicted in figure below.

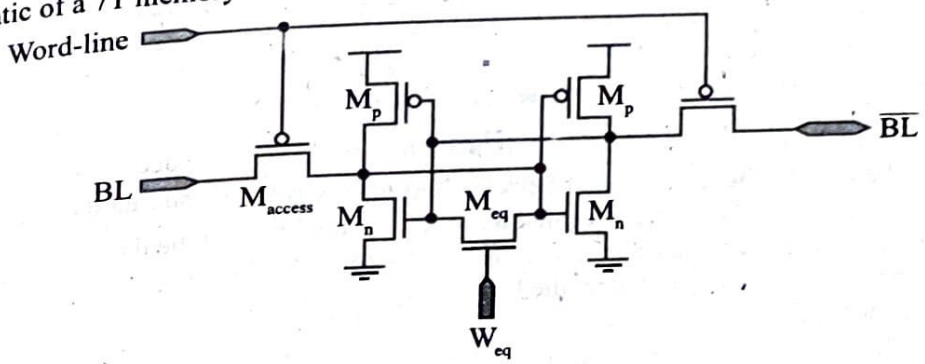


Figure: 7T Memory Cell

It can be viewed as a 6T memory cell added with one more transistor Meq for equalization. Before each write operation, Meq is enabled to delete the contents of the cell. It is then turned off during reading to avoid interfering with the normal reading process. The current conveyor can detect a small differential current from the memory cell during read access since both read and write operations are performed in current-mode. Similar to a small differential data current, a memory cell's contents can be altered. A small differential current can be well conducted by pMOS devices, making them suitable for use as access transistors.

Q30. Explain loadless CMOS 4T SRAM cell.

Model Paper-1, Q11(a)

Ans:

A loadless CMOS 4T SRAM cell is developed for ultra-high density SRAM macros and is depicted in figure (1).

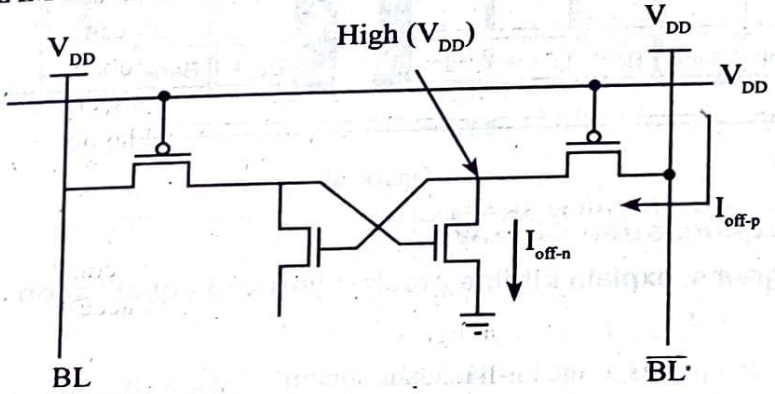


Figure (1): Loadless CMOS 4T SRAM Cell

It consists of two minimal-size NMOS transistors that acts as drive transistors and two minimal-size PMOS transistors that acts as access transistors.

This configuration enables the two cell nodes to store a full-swing signal during a write operation.

During stand-by mode, stable data retention is achieved by precharging bit-lines with the supply voltage (V_{DD}), while access transistors are turned off and repurposed as load components.

Inorder to ensure stable data retention and to account for the Ioff process distribution, the off-state current of the pMOS (I_{off-p}) must be larger than that of the nMOS (I_{off-n}) in order to maintain the data without a refresh cycle.

The 4T SRAM cell occupies 50-65% area of the area of a conventional 6T SRAM cell. So it is an area-efficient device.

Figure (2) depicts a 1T SRAM bit cell with an access transistor and a capacitor. It is essentially a planar DRAM with a MOS structure for the storage capacitor.

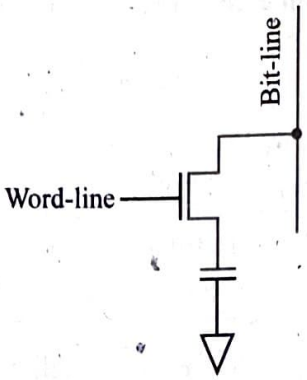


Figure (2)

To avoid data loss, the data in these cells must be refreshed on a regular basis. The data is stored in the form of charge in the cell. The charge held in the capacitor is transformed to voltage at the bit-lines during a read operation, and then amplified by a sensing amplifier. The bit-line voltage has a major impact on memory access speed. Because the bit-line voltage is related to the ratio of cell capacitance to bit-line capacitance, the bit-lines must be kept short to ensure modest cell size. As a result, the bit-line capacitance will be reduced.

In addition to improving access speed, the short bit-line and word-line structure may be easily structured to create several small independent memory banks, as shown in figure (3). The multibank architecture allows for simultaneous functions in each bank, allowing for a hidden refresh cycle.

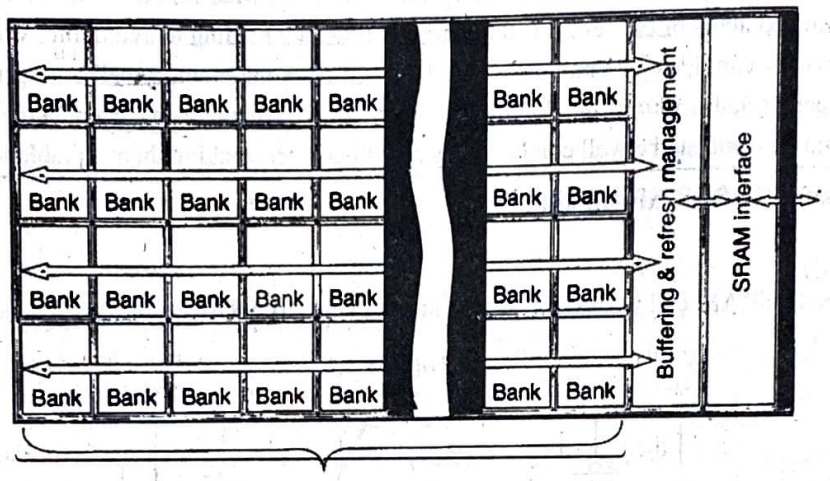


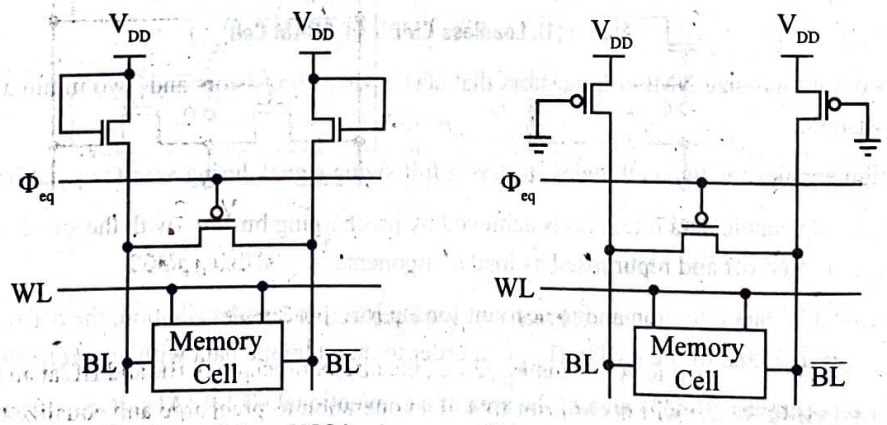
Figure (3)

5.2.3 Precharge and Equalization Circuit

Q31. With neat circuit diagrams, explain bit-line precharging and equalization in a static RAM.

Ans: Model Paper-2, Q11(a)

Figure (1) shows nMOS and pMOS as the bit-line loads conditioning circuits .



(a) Bit-line Precharge Using nMOS loads (b) Bit-line Precharge Using pMOS Loads

Figure (1)

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

MOS as the Bit-line load Conditioning Circuit

nMOS bit-line loads are mostly used in older SRAM applications that precharged the bit-lines to $V_{DD} - V_T$ with a 5V ower supply voltage. For voltage-mode sense amplifiers with a high gain and low sensing delay, the precharged voltage level adequate. However, that the precharge level will be lower than 2V if V_{DD} is reduced(scaled down) to 3 V or less. As a result, uring a read operation, the memory cell's noise margin is reduced significantly, resulting in decreased cell stability. As a result, e conditioning circuit with nMOS as the load becomes inappropriate.

MOS as the Bit-line load Conditioning Circuit

When the precharge level will be lower than 2 V if V_{DD} is reduced(scaled down) to 3 V or less, as a result, the precharged it-line voltage is V_{DD}. So, a pMOS device is required since the V_{DD} bit-line load is less than 3 V.

At the end of each memory cycle, there is a difference voltage on the bit lines. From the waveforms in figure (2), to il- strate the need for equalization, it can be seen that restoring the differential voltages prior to use will result in a faster read time smaller delay time Δt).

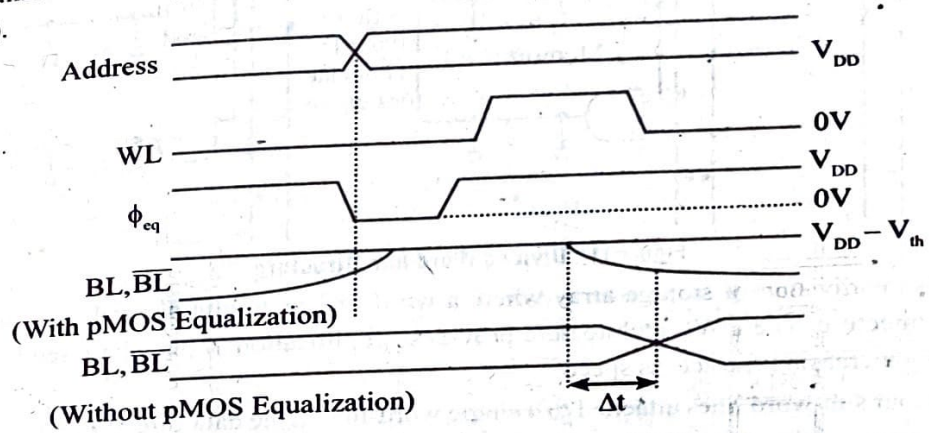


Figure (2): Bit-line Equalizing and Conditioning Waveforms

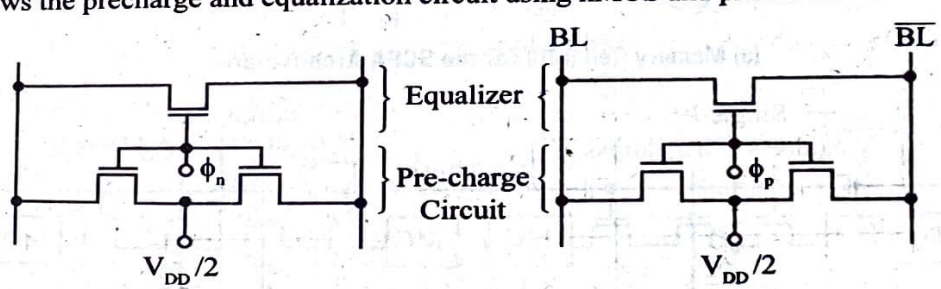
A pMOS device performs this operation, which is triggered by an Address Transition Detection (ATD) circuit. When the ddress is changed, the ATD circuit emits an equalizing falling pulse ϕ_{eq}. As a result, the bit lines become equalized after each ead and write.

Transmission gates are used in the column decoder because they speed up the transfer of data from the bit lines to the ommon data lines.

Q32. Explain bit-line precharging and equalization in SRAM using MOSFET.

Ans:

Figure below shows the precharge and equalization circuit using nMOS and pMOS FETs as the bit-line loads.



(a) Using nMOS **(b) Using pMOS**

Figure: Precharge and Equalization Circuit

In this circuit, the precharged bit-line voltage is $V_{DD}/2$ i.e., the circuit precharges BL and $\overline{BL}$ at an intermediate value typi- ally at $V_{DD}/2$. A clock signal (ϕ_n or ϕ_p) will occur prior to a read operation to precharge and equalize the bit-line capacitances o $V_{DD}/2$. This step is critical in ensuring that the sense amplifier picks up the correct signals from the memory cells. Precharge nd equalization circuits will also reduce access time.

5.2.4 Low-Power SRAM Technologies

Q33. Explain the divided word line architecture of SRAM.

Ans:

When the storage capacity of a RAM increases, the effect of parasitic resistance and capacitance associated with the interconnect lines also increases. To reduce these effects, multiple storage cores are employed which partitions the total storage array into separate blocks of cells, where each block represents a particular range of address. This partitioning is continued to form sub-sub blocks as shown in figure (1).

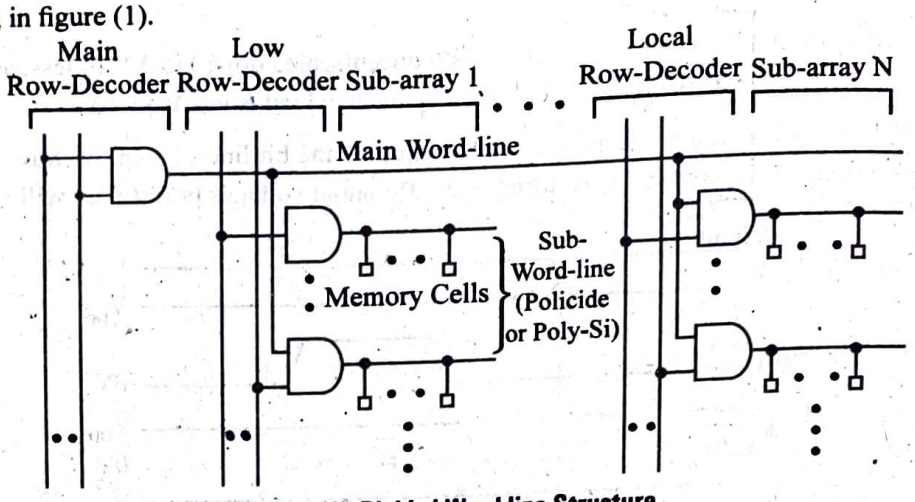
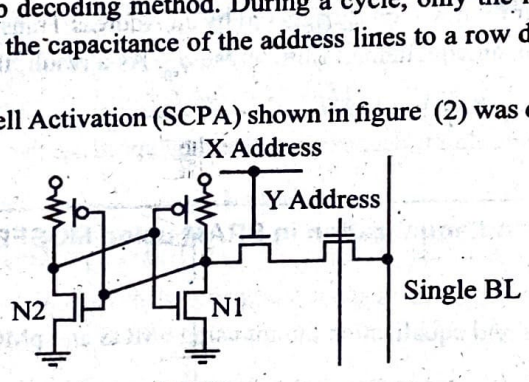


Figure (1): Divided Word-line Structure

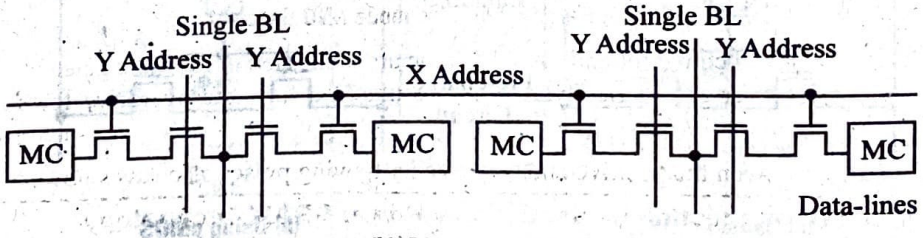
Figure (1) describes the division of storage array where a word line routing is divided into multiple paths using DWL (Divided Word - Line) architecture. The DWL architecture provides simplification of decoder circuits and distributes the loads over multiple stages thereby increasing the access speed.

There are normally four sub-word lines attached to a single word-line in the data-direction. The main row decoder and the local row decoder are combined in this design. The main word line and the associated local row decoder choose the sub-word line of each subarray using the two-step decoding method. During a cycle, only the memory cells in the specified subarray's sub-word-line are accessible. To reduce the capacitance of the address lines to a row decoder and the word-line RC delay, it is recommended to use this method.

The Single bit-line Cross-Point cell Activation (SCPA) shown in figure (2) was developed to further reduce power usage.



(a) Memory Cell (MC) for the SCPA Architecture



(b) Diagram of SCPA

Figure (2)

Both the access transistors and the X address are controlled by the Y address. A column current is only drawn by the accessible memory cell since only one memory cell is activated at the cross point of *X* and *Y*. Thus, this technique needs the smallest possible column current without increasing the cell array's block division. Because of this, the decoder is smaller and the memory core is less than in a traditional DWL scheme by 10%.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

4. Explain AC reduction technique used in low-Power SRAM Technology.

ns:

In low-power SRAM applications, the AC current is reduced usually using the two-stage decoding architecture depicted figure below.

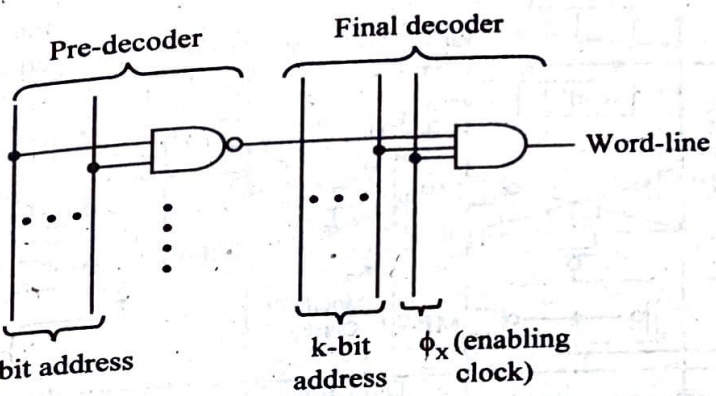


Figure: Two-stage Decoding Architecture

In this architecture, the number of transistors, fan-in and the loading on the address input buffers are decreased when mpared to a single-stage decoder. It increases and optimizes speed and power characteristics.

5. Explain pulse operation technique used in low-Power SRAM Technology.

ns:

The pulse operation technique in low-power SRAM technology is used to,

Reduce the active duty cycle duration

Reduce power dissipation.

Lowering the signal swing on high capacitance predecode lines, write-bus lines, and bit-lines can minimize power nsumption during write and decode operations without compromising performance. Usually, the precharging is done at DD/2. It greatly saves power but at the same time, decreases gate overdrive at the receiver end. A method using the half-swing lse-mode scheme is introduced to overcome this problem as depicted in figure below.

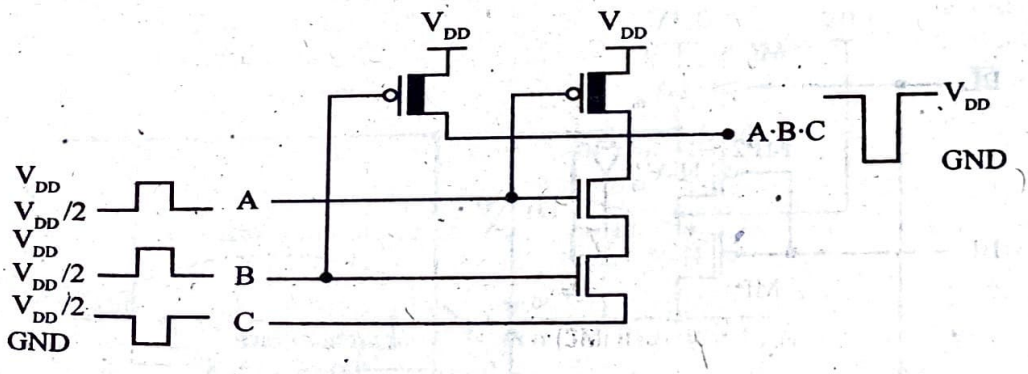


Figure: Half-swing Pulse-mode AND Gate

In this technique, a voltage-level converter and an AND logic are combined to form the gate. The positive and a negative lf-swings are added with the receiver-gate logic style, which results in a full gate overdrive for all transistors driving the rward transition.

Further, charge recycling between the positive and negative half-swing pulses also saves power.

6. Explain low-power sensing technique used in low-Power SRAM Technology.

ns:

In low-power sensing technique, the current mirror voltage circuit in replaced with a current-mode sense amplifier in sense circuit: The current-mode sense amplifier provides very low bit-line and data-line voltage swings and allows for low- oltage and high-speed operation.

Figure below illustrates an enhanced current mode sense amplifier that was built with a modified current conveyor for 1.5 V wer supply.

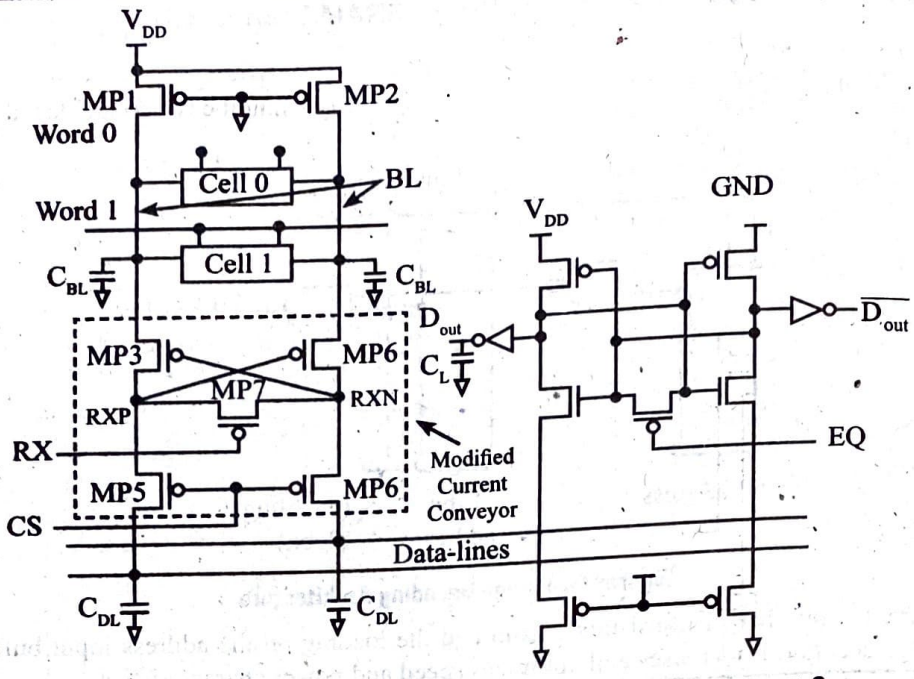


Figure: An Improved Current-mode Sense Amplifier with a Modified Current Conveyor

The advantage of this enhanced circuit is solves the problem of pattern dependency, which restricts the scaling of the operating voltage. MP7 pMOS transistor is added to the modified current conveyor to make it run faster. To eliminate any residual voltage differences between RXP and RXN after each read cycle, an equalization transistor will be turned on.

Q37. Explain leakage current technique used in low-Power SRAM Technology.

Ans:

Dynamic power consumption can be reduced by lowering the operational power supply voltage, V_{DD}. However the threshold voltage will also reduce on scaling V_{DD}. As a result, the leakage current increases in both active and standby modes of operation. This problem has can be overcome through a variety of methods. One such is the Auto Backgate Controlled Multi-Threshold Voltages (ABC MT-CMOS) technique as depicted in figure below.

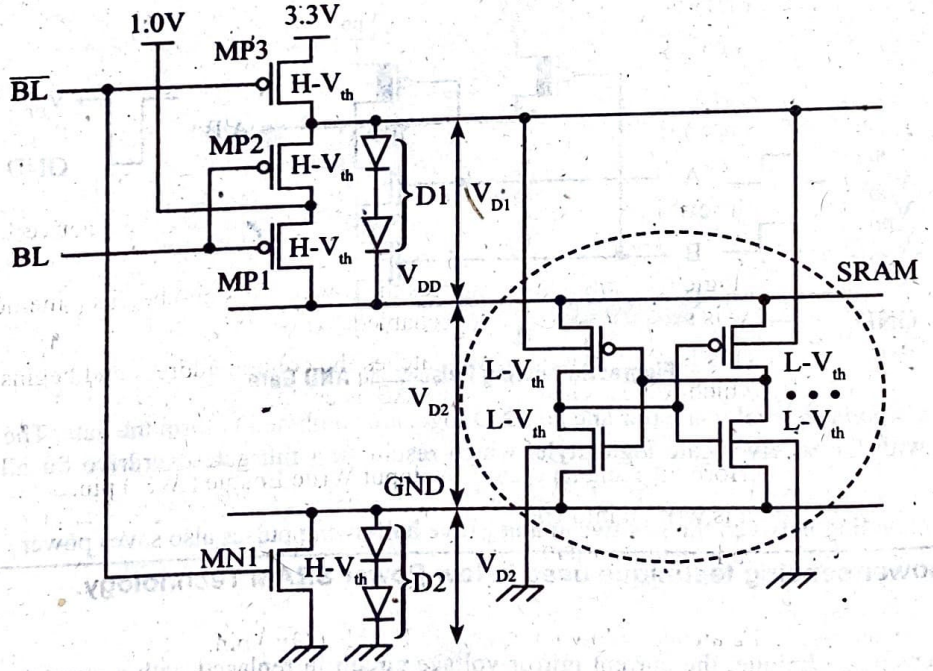


Figure: Schematic Diagram of the ABC MT-CMOS

Low-V_{th} (L-V_{th}) devices are used in the internal circuit design. MP1, MP2, MP3, and MN1 are external transistors with high voltage thresholds (H-V_{th}) that are used to cut off the leakage current. This technique can significantly reduce the leakage current in sleep-mode.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

5.3.1 Basics of DRAM

Q38. Draw the block diagram of a dynamic RAM architecture and explain its operation.

Ans: The block diagram of a 4 Mb Dynamic RAM(DRAM) architecture is depicted in figure below.

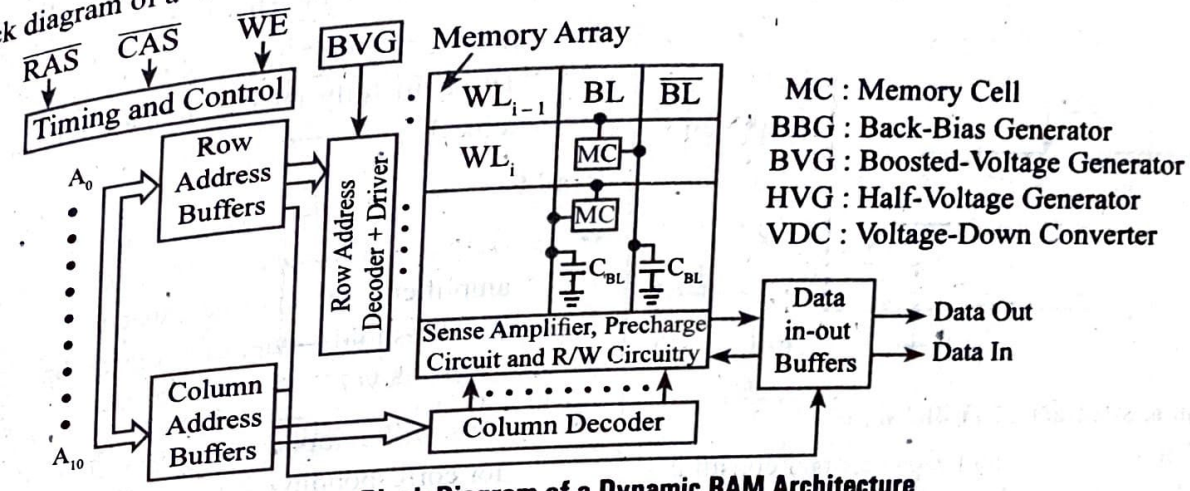


Figure: Block Diagram of a Dynamic RAM Architecture

The major blocks of the DRAM are,

1. Memory cells (MC)
2. Row and column decoders
3. Word-line drivers
4. Sense amplifiers
5. Precharge and equalization circuits
6. Read/write circuitries
7. Half-Voltage Generator (HVG)
8. Back-Bias Generator (BBG)
9. Boosted-Voltage Generator (BVG)
10. Voltage-Down Converter (VDC)
11. In/Out (I/O) buffers.

The following select/control pins are used in the operation of DRAM.

Address select ($\overline{AS}$)

The address logic is used for both row and column addresses in time-slots. So it has two active-low select/control lines $\overline{RAS}$ and $\overline{CAS}$.

If Row Address select ($\overline{RAS}$) is logic 0 or low, then it latches the row address and begins a memory cycle. The address pin data will determine which row is selected when RAS is enabled.

If Column Address select ($\overline{CAS}$) is logic 0 or low, then it latches the column address and begins a memory cycle. The address pin data will determine which row is selected when $\overline{CAS}$ is enabled.

Write Enable ($\overline{WE}$)

Read and write operations are performed using an active low input Write Enable ($\overline{WE}$) pin.

When $\overline{WE} = 0$ or logic low, data is written into DRAM

When $\overline{WE} = 1$ or logic high, data is read from DRAM.

Output Enable ($\overline{OE}$)

Data buffering is carried out using an active low input Output Enable ($\overline{OE}$) pin.

When $\overline{OE} = 0$ or logic low, signals allow data to enter the data input/output buffer during a read operation.

DRAM During a write process, $\overline{OE}$ is deactivated.

Data Input/Output (I/O)

The data input/output pins are used to connect the DRAM to the rest of the system.

Q39. Draw the structure of a 1T DRAM architecture and explain its operation.

Model Paper-1, Q11(b)

Ans:

The circuit diagram for a one-transistor dynamic RAM core cell is shown in figure.

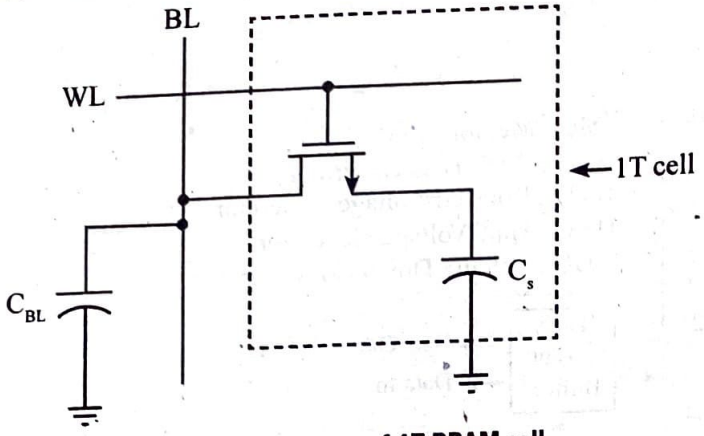


Figure: Structure of 1T DRAM cell

- A single nMOS transistor and capacitor pair constitutes each DRAM memory cell.
- From the figure, it is noted that the cell consists of two external connections which are Bit-Line(BL) and Word-Line(WL).
- The value is stored on a capacitor guarded by a transistor as shown in the figure.
- Word line (WL) is the row decoder output signal and controls the transistor gate that runs perpendicular to the Bit-Line .
- By setting the word line to high, the capacitor connects to the bit line.
- The bit line is set according to a new value written and the capacitor is forced to the proper value.
- While reading the value, the bit line is first precharged before the word line is activated.
- When the storage capacitor is discharged, the charge will flow from the bit line to the capacitor by lowering the voltage on the bit line.
- The capacitance C_{BL} represents the capacitance of BL and parasitic capacitance.
- The capacitance C_S represents the storage capacitor. Data is stored in the form of charge in it.
- This cell's value must be refreshed periodically, which can be done by reading the cell.

Q40. Explain the read and write operations in DRAM.

Ans:

To store the data in the cell, the charge must be stored in the capacitor. The bit-line parasitic capacitance (C_{BL}) is precharged to logic half-V_{DD} before the read operation. After that, WL is pulled high to turn on the nMOS-type access transistor. The row and column values determine which memory cell to read from in an array of memory cells.

Read 1

- When the data stored in the selected cell is to be read, its transistor will be turned ON.
- Then, the sense amplifier which is connected to the bit line, checks whether the capacitor's charge (C_s) is more than the threshold value.
- If it is, then it will provide full voltage to the bit line. This will fully charge the capacitor corresponding to logic 1.

Read 0

- Here, C_{BL} discharges and C_s charges up. If the sense amplifier finds that the capacitor's charge is less than the threshold value, then the bit line will be pulled to ground level.
- This will completely remove the charge of the capacitor corresponding to logic '0'.

Write 1

- By Charging C_{BL} to either V_{DD} or ($V_{DD} - V_{th}$) based on the type of write circuitry employed, the voltage at BL is made to logic 1, while a row address decoder pulls the appropriate word line high.
- The nMOS transistor's threshold voltage is referred to as V_{th}.
- By asserting the WL, the access transistor will conduct and charge the storage capacitor to a logic "1" level.

Write 0

- The write circuitry pulls the bit-line to logic 0, while the row address decoder pulls-up the appropriate word-line.
- If the access transistor is initially storing a logic 1, the storage capacitor discharges; if it is already storing a logic 1, the storage capacitor remains at logic 0.

5.3.2 Self-Refresh Circuit

Points to Remeber
❖ Capacitors are used in DRAM to store charges, however these charges tend to deteriorate over time. As a result, before the stored voltages reach intolerable levels, the DRAM must be refreshed to its original state.
❖ Each memory cell in a selected row is automatically refreshed when you read or write, but the whole memory array needs to be refreshed every 5 to 10 ms, depending on the chip.

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

Q41. With a block diagram, explain the general mechanism of a self-refresh circuit in DRAM.

Model Paper-2, Q11(b)

Ans:

Memory array refresh operations need the usage of a circuitry capable of providing a refreshing cycle. Simultaneously, this circuit must be able to identify and isolate the row that requires refreshing, allowing the DRAMs to get a timely refresh. Refreshing will continue until all of the memory cells have been properly refreshed once it has begun.

The refresh circuit can be implemented in following two ways,

1. Using an address decoder and a counter. Starting from 0, the counter will count up to the maximum amount of entries possible.
2. Using a shift register. It can be used to choose only one row at a time, and then shift the selection to the next row in a sequential manner. This method requires clock, reset and refresh trigger pins.

The block diagram of the refresh trigger signal generator is as shown in figure below.

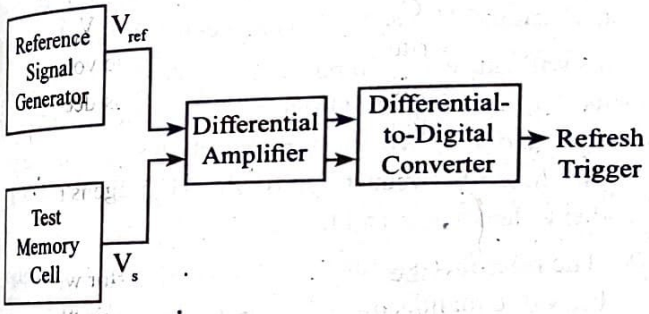


Figure: Block Diagram of the Refresh Trigger Signal Generator

The main components of the refresh trigger signal generator are,

1. Test memory cell
2. Reference signal generator
3. Differential amplifier
4. Differential to digital converter

Test Memory Cell

It serves as a test bed for monitoring decreasing stored data (or charge), V_s.

V_s represents a logic "1" that degrades with time as measured by the differential amplifiers.

Reference Signal Generator

It generates a V_{ref} reference signal that is compared to V_s. The signal V_{ref} is held constant at 2V, which is the lowest permissible value for the stored "1" in the memory array to decay to.

Differential Amplifier

The nMOS differential amplifier compares the voltages V_s and V_{ref} then outputs the difference signal.

4. Differential to Digital Converter

- The CMOS differential to digital converter then converts the difference signal from differential amplifier into a single-ended output which serves as the refresh trigger signal to start the refresh cycle.

The refresh trigger signal operates as a control pin that is only active when the refresh operation is required.

Q42. With neat circuit diagram, explain the operation of,

1. Reference signal generator

2. Test memory cell.

Ans:

1. Reference signal generator

The circuit schematic of a reference signal generator is illustrated in figure (1). It has four transistors , T_w, T_c, T_1 and T_r along with a write line.

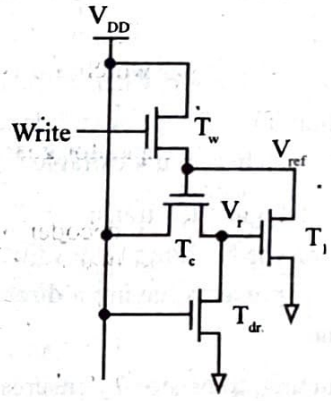


Figure (1): Circuit Schematic of a Reference Signal Generator

- When the write-line is at logic high, the transistor, T_w is used to transmit a logic 1 to the storage device.
- The gate of the transistor, T_c acts as the storage element which enables a logic 1 to be transmitted to node V_r after a logic 1 has been written to V_{ref}.
- The transistor, T_1 serves as a temporary channel for leaking from V_{ref} to the ground. It reduces V_{ref} from $V_{DD} - V_{th}$ to a smallest value such that the stored logic 1 is decayed.
- The transistor, T_{dr} is always on inorder to drain the node V_r and turn-off T_1.
- At periodic intervals, the write-line signal is pulled to logic 1 so that a "1" can be passed from the write-line to the storage node.

2. Test Memory Cell

The circuit schematic of a test memory cell is illustrated in figure (2). It has four transistors T_{w1}, T_{w2}, T_{c1} and T_{rx}. These transistors are connected to write line, bit line and read line.

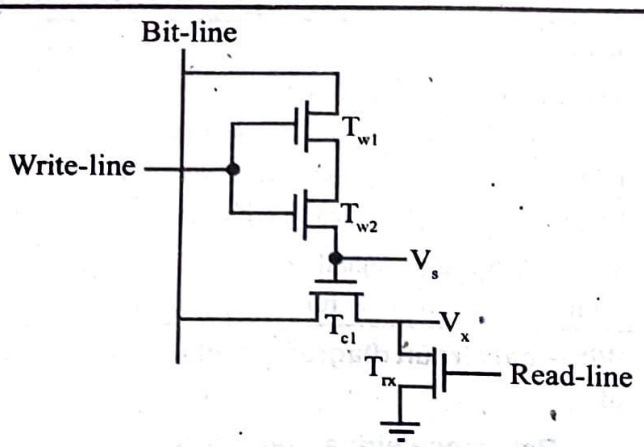


Figure (2): Circuit Schematic of a Test Memory Cell

- Data is passed from the bit-line to node V_s through serial transistors T_{w1} and T_{w2}.
- The voltage passed by this series combination is slightly lower than the voltage passed by a memory cell with only one transistor. As a result, this voltage will drop more quickly than a "1" stored in a standard memory array.
- It creates a safety margin when refreshing the stored data, such that that the refresh happens before the stored voltage reaches an undesirable level.
- When V_s is at logic "1," transistor T_{c1} compares the logic values on the bit-line to the value stored at node V_s, whereas T_{rx} prevents having a direct link from bit-line to ground.
- When conducting, transistor T_{rx} ensures that node V_x is at logic "0."
- If the write signal is at logic "0," leakage slowly decays the stored "1" at V_s.
- After logic "1" is written to V_s, the information on the bit-line is modified frequently to indicate the data on the bit-lines of memory arrays.

The voltages V_s from the test memory cell and V_{ref} generated by and the RSG will be sent to the nMOS differential amplifier in the next stage.

Q43. With neat circuit diagram, explain the operation of ,

1. The nMOS differential amplifier

2. CMOS differential to digital converter.

Ans:

1. The nMOS Differential Amplifier

The circuit schematic of the nMOS differential amplifier is shown in figure (1). It has five transistors T_{L1}, T_{L2}, T_{ref}, T_s and T_{CS}.

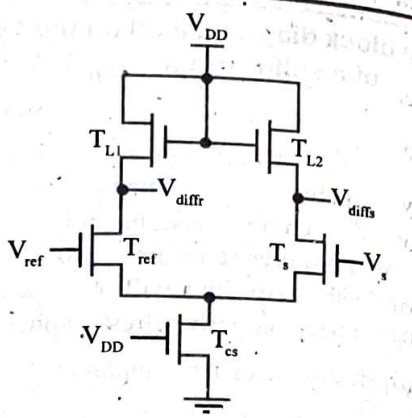


Figure (1): Circuit Schematic of a nMOS Differential Amplifier

- The resistive active loads for this active amplifier are the transistors T_{L1} and T_{L2}. These transistors are always on.
- The reference signal V_{ref} drives transistor T_{ref}, whereas the decaying stored signal V_s drives transistor T_s. The transistors T_{ref}, T_{L1}, and T_{cs} are always conducting, hence, there is always a direct path from V_{DD} to ground.
- The voltage difference between V_{diffr} and V_{diffs} is at its maximum level when V_s is at a constant logic "1."
- Ts will cause V_{diffs} to be discharged to logic "0" as long as V_s is an acceptable logic "1." The voltages values of V_{diffs} and V_{diffr} will become nearly equal resulting the difference between to zero when V_s falls to equal V_{ref} due to leakage current.
- The reference signal V_{ref} represents the lowest admissible value that the stored value V_s is allowed to decrease to.

The two outputs V_{diffs} and V_{diffr} are fed as inputs to the next the CMOS differential-to-digital converter.

2. CMOS Differential to Digital Converter

The circuit schematic of a CMOS differential to digital converter is illustrated in figure (2). It has five transistors T_{cm1}, T_{cm2}, T_{pd1}, T_{pd2} and T_{CS} along with an inverter at the output.

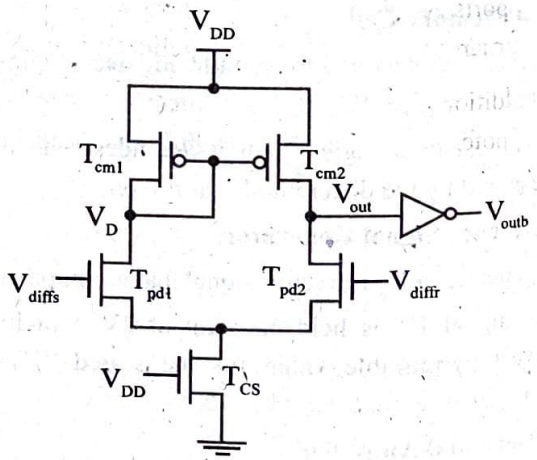


Figure (2): Circuit Schematic of a CMOS Differential to Digital Converter

WARNING: Xerox/Photocopying of this book is a CRIMINAL act. Anyone found guilty is LIABLE to face LEGAL proceedings.

It has a current-mirror formed by the transistors , T_{cm1} and T_{cm2}. The gate terminals of T_{cm1} and T_{cm2} are connected to the drain terminal of T_{cm1} as shown in figure.

The transistor T_{pd2} is always on.

Initially the voltage V_{diffs} at transistor T_{pd1} is 0. Hence, the voltage V_D does not decay to 0. So the transistors T_{cm1} and T_{cm2} does not turn on.

As V_{diffs} increases gradually to logic1, the transistor T_{pd1} turns on and starts conducting whereas the voltage V_D starts decaying to logic 0. Hence, the T_{cm1} and T_{cm2} turn on. As a result, the node V_D is raised to logic 1 from logic 0.

The inverter complements the output V_{out} to V_{outb} which acts a refresh trigger signal. This signal is used to start refresh cycle.

3.3 Future Trend and Development of DRAM

4. Explain future trend and development of DRAM.

Model Paper-3, Q11(b)

s:

The standard DRAM has made good advances, with memory-chip capacity increasing from 4 Gb in R&D and 256 Mb mass production. Furthermore, the new memory subsystem method, 1.6 GByte/s DRAM with Flexible Mapping Redundancy hnique and Additional Refresh Scheme, has increased DRAM throughput to 1.6 Gbyte/s for a 1 Gb chip.

Then DRAM entered multigigabit era. This advancement has been aided by high density semiconductor technologies like he one-transistor one-capacitor (1T) cell, vertical cell capacitors, scaled CMOS devices, and multilevel metal wiring, which e quadrupled memory capacity. However, as chip sizes grow larger, it becomes more difficult to achieve faster speeds be-se, as memory capacity increases, both wiring capacitance and wiring resistance increase as well. Future speed enhancements l have to rely on true breakthroughs and innovations in high-performance semiconductor circuit design.

Several DRAM architectures have the potential in terms of performance and are highly recommended for future devel-nent. The Rambus DRAM compete directly with the industry-standard synchronous DRAM, SyncLink DRAM, and DDR AM architectures.

The SLDRAM Consortium is a nonprofit organization that promotes the development of an open standard for the DRAM. It was founded in 1998. It rivals Rambus by bringing together engineering talent from most of the DRAM and system s to focus on the challenge of meeting high bandwidth DRAM requirements.

. The SLDRAM technology is an advancement to SDRAM. It has the following features.

SLDRAM uses a multiplexed command bus with fewer pins

Obtains faster bandwidth via external speeds.

It supports many separate banks for video, graphics, and telecommunications applications, enabling for quick read/write bus turnaround and tiny, fully pipelined bursts.

In addition, in the DRAM architecture competition, the EmDRAM for system-on-chip (SOC) applications has shown endous potential. Rather of focusing on cheap cost and low voltage, EmDRAM has recently emphasized high-speed tioning.

IMPORTANT QUESTIONS

Q1. Draw and explain the architecture of a basic ROM.

Ans: Refer Q16.

Important Quest

Q2. With neat circuit diagrams, explain the basic structures of ROM cell core using nMOS transistors.

Ans: Refer Q17.

Important Quest

Q3. What are the low-power techniques used at architecture level of ROM? Explain.

Ans: Refer Q19.

Important Quest

Q4. What are the low-power techniques used at circuit level of ROM? Explain.

Ans: Refer Q20.

Important Quest

Q5. Explain future trend and development of ROM.

Ans: Refer Q21.

Important Quest

Q6. With neat block diagram, explain the architecture of a static RAM(SRAM).

Ans: Refer Q22.

Important Quest

Q7. What is a memory cell? Explain the basic 6T SRAM cell consist of cross coupled CMOS inverters.

Ans: Refer Q24.

Important Quest

Q8. Explain the racing between 6T SRAM and 4T SRAM memory cells.

Ans: Refer Q26.

Important Quest

Q9. Distinguish between 6T SRAM cell and 4T SRAM cell.

Ans: Refer Q27.

Important Quest

Q10. With neat diagram, explain any one design method to achieve low-voltage low-power SRAM cell.

Ans: Refer Q28.

Important Quest

Q11. Explain loadless CMOS 4T SRAM cell.

Ans: Refer Q30.

Important Quest

Q12. With neat circuit diagrams, explain bit-line precharging and equalization in a static RAM.

Ans: Refer Q31.

Important Quest

Q13. Explain the divided word line architecture of SRAM.

Ans: Refer Q33.

Important Quest

Q14. Explain pulse operation technique used in low-Power SRAM Technology.

Ans: Refer Q35.

Important Questio

Q15. Draw the block diagram of a dynamic RAM architecture and explain its operation.

Ans: Refer Q38.

Important Questio

Q16. With a block diagram, explain the general mec-hanism of a self-refresh circuit in DRAM.

Ans: Refer Q41.

Important Questio

proceedings.